UNDERSTANDING AND MODELING FOOD FLOW NETWORKS ACROSS SPATIAL SCALES

BY

XIAOWEN LIN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Assistant Professor Megan Konar, Chair
Professor Murugesu Sivapalan
Professor Praveen Kumar
Assistant Professor Kaiyu Guan
Associate Professor Benjamin L. Ruddell, Northern Arizona University

ABSTRACT

We live in an increasingly global society, in which food commodity transfers enable production and consumption activities to be separated in space via complex supply chains. Here, we refer to the movement of food commodities from one location to another as 'food flows', reserving the term 'food trade' for the international exchange of food commodities between countries. Food flows underpin the complex food supply chains that are prevalent in our increasingly globalized world. Recently, much effort has been devoted to evaluating the resources (e.g. water, carbon, nutrients) embodied in food trade. Now, research is needed to understand the scientific principles of the food commodity flows that underpin these virtual resource transfers. What are the network properties of food flows within a country? How do food flows vary with spatial scale? How can we model food flows in locations without empirical information? This dissertation seeks to address these three overarching questions.

First, this dissertation presents a novel application of network analysis to empirical information of domestic food flows within the USA, a country with global importance as a major agricultural producer and trade power. We find normal node degree distributions and Weibull node strength and betweenness centrality distributions. An unassortative network structure with high clustering coefficients exists. These network properties indicate that the USA food flow network is highly social and well-mixed. However, a power law relationship between node betweenness centrality and node degree indicates potential network vulnerability to the disturbance of key nodes. We perform an equality analysis which serves as a benchmark for global food trade, where the Gini coefficient = 0.579, Lorenz asymmetry coefficient = 0.966, and Hoover index = 0.442. These findings shed insight into trade network scaling and proxy free trade and equitable network architectures.

Second, this dissertation presents an empirical analysis of food commodity flow networks across the full spectrum of spatial scales: global, national, and village. We discover properties of both scale invariance and scale dependence in food flow networks. The statistical distribution of node connectivity and mass flux are consistent across scales. Node connectivity follows a generalized exponential distribution, while node mass flux follows a Gamma distribution across scales. Similarly, the relationship between node connectivity and mass

flux follows a power law across scales. However, the parameters of the distributions change with spatial scale. Mean node connectivity and mass flux increase with increasing scale. A core group of nodes exists at all scales, but node centrality increases as the spatial scale decreases, indicating that some households are more critical to village food exchanges than countries are to global trade. Remarkably, the structural network properties of food flows are consistent across spatial scales, indicating that a universal mechanism may underpin food exchange systems.

Finally, we use our understanding of food flow networks across spatial scales to model food flows at resolutions for which empirical information is not available. Detailed spatial information on food flows is rare, but it is increasingly important to understand spatially resolved food flows to assess their embodied resources and vulnerability to supply chain disturbances. To this end, we develop the Food Flow Model, a data-driven methodology to estimate spatially explicit food flows for subnational locations without data. The Food Flow Model integrates machine learning, network properties, production and consumption statistics, mass balance constraints, and linear programming. We use the Food Flow Model to infer food flows between counties within the United States. Specifically, we downscale empirical information on food flows between 132 Freight Analysis Framework (FAF) locations (17,292 potential links) to the 3,142 counties and county-equivalents of the United States (9,869,022 potential links). Future work can build on these efforts to improve our understanding of vulnerabilities within a national food supply chain, determine critical infrastructures, and enable spatially detailed footprint assessments.

ACKNOWLEDGMENTS

It is the support of many people that makes this work possible. I would like to express my gratitude to my advisor Dr. Megan Konar for her guidance, patience, expertise and all the support she has provided, that a graduate student could ever found from the best advisor. I would also like to thank my lab-mates, Qian Dang, Landon Marston, Nicole Jackson, and Paul Ruess, for their collaboration and inspirations. Last but not least, I would like to thank my committee members besides my advisor, including Murugesu Sivapalan, Praveen Kumar, Kaiyu Guan, and Benjamin R. Ruddell, who have offered indispensable insights and suggestions.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation and overview

Commodity production and consumption activities are separated in space. This separation is enabled by the specialization of production in certain regions and through globalization, which has made transportation cheaper and reduced barriers to commodity exchange. According to *Han and Soroka* (2014), the majority of commodities are now transferred elsewhere to be consumed. In this dissertation, the movement of commodities from one location to another is referred to as "commodity flow". This terminology is adopted to distinguish commodity flows from "commodity trades", which refer to the specific case of commodity exchange between nations. Here, the focus is on "food flows", or the movement of food commodities between spatial units, which may be countries in international trade, locations within a country, or households within a village.

Food flows are inextricably linked to water resources and security (*Aldaya et al.*, 2012). Food flows are often enabled by the ability to use water resources in agricultural production and food processing activities. Often times, agricultural production and food flows are enabled by using groundwater from critical national reserves (*Marston et al.*, 2015). For example, California is known as the "fruit and vegetable basket" of the United States. California is able to maintain this status by over-exploiting its groundwater to produce fruits, vegetables and nuts, especially during drought (*Marston and Konar*, 2017). Water deficient regions depleting local water resource to produce and exchange agricultural and food commodities occurs around the world (*Berkoff*, 2003; *Micklin*, 1988; *Salako and Tian*, 2003). In this way, food flows and trade can be thought of as the exchange of the water embodied in those commodities, or its "virtual water" flows and trade (*Konar et al.*, 2016a).

Food flows are also a major factor reshaping the landscape of food security. According to (*Foundation*, 2018), 1/7 of the world population is hungry, while 1/3 of food is wasted despite sufficient global production. These statistics emphasize the inequality in distribution that leads to food security issues, instead of deficiency in production. While chemical and agricultural companies are investing billions into discovering solutions towards higher food

productivity, food distribution has been largely neglected. Nevertheless, due to economic development, population growth, and climate change, inequalities in the distribution of food has been accelerating over the last several decades (*Godfray et al.*, 2010; *Kastner et al.*, 2012). Understanding the structure of food flow networks in both space and time can help us to better understand inequalities in food distributions and opportunities for improvements. Recent work suggests that reducing trade frictions will reduce poverty and save water under a changing climate (*Hertel et al.*, 2010; *Konar et al.*, 2016b).

Information on food flows at the spatial unit smaller than the nation is sparse. Excellent data on international trade is available. This data is available annually from the 1960s onwards and provides detailed commodity information. Conversely, little is known about how food commodities are moved around within a single nation, within regions, or within villages/cities. Since these food supply chains connect consumers with producers and may be vulnerable to supply chain shocks going forward, it is increasingly important to evaluate their spatial structure. This makes it increasingly important to estimate food flows in locations that do not have empirical data available. Then, equipped with this understanding, future work on the network structure of these food flows and their interconnections with other infrastructure networks can be used to shed light on their vulnerabilities and resiliencies.

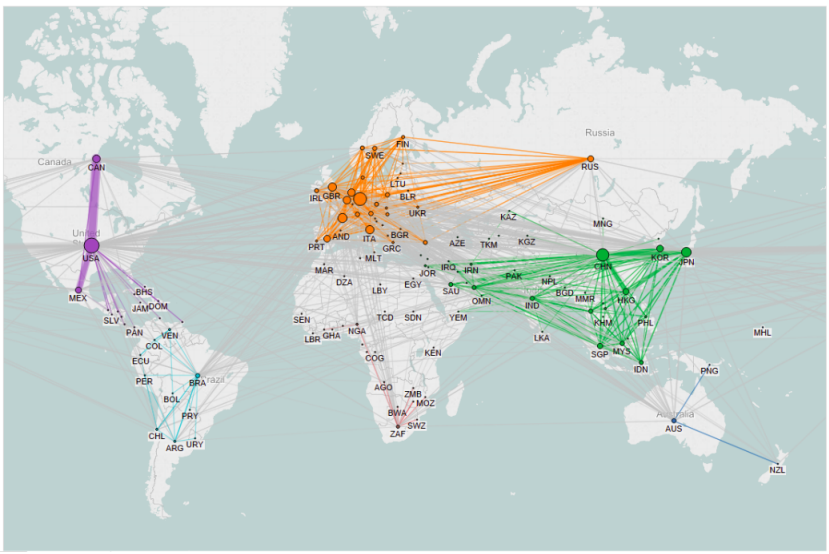

Figure 1.1: Country Network

1.2 Background

Network analysis has been increasingly used to understand complex systems. This recent interest in complex networks is largely due to the discovery of organizing principles in networks (*Newman et al.*, 2006; *Costa et al.*, 2007), such as community structure (*Watts*

and Strogatz, 1998) and scale-free properties (*Barabási and Albert*, 1997). Its flexibility and ability to represent real-world systems has been well demonstrated across multiple areas (*Barabasi*, 2002), including transpiration systems (*Kalapala et al.*, 2006; *Masucci et al.*, 2009; *Kaluza et al.*, 2010), the world wide web (*Barabási and Albert*, 1997), international tourism (*Miguens and Mendes*, 2008), financial transactions (*Garlaschelli and Loffredo*, 2005; *Kyriakopoulos et al.*, 2009), and scientific collaborations (*Newman*, 2001a; *Barrat et al.*, 2004), among others. Specifically, network analysis and theory has been demonstrated effective in commodity flow study in recent studies (*Serrano and Boguna*, 2003; *Garlaschelli and Loffredo*, 2005; *Bhattacharya et al.*, 2007; *Fagiolo et al.*, 2008; *Barigozzi et al.*, 2010). Recent work has begun to focus on the network characterization of global food trade and flow (*Konar et al.*, 2011; *Ercsey-Ravasz et al.*, 2012; *Shutters and Muneepeerakul*, 2012). Adopting similar tools and analysis framework will enable us to directly compare with these new development and build on top of this knowledge base.

In order to understand food flow network generation mechanism, network analysis alone won't be sufficient. Furthermore, statistical network modeling is utilized to explain and understand network characteristics, both topology and link strength. Enormous literature has studied modeling of network formation through different approaches, including static random graph models of network, growing random networks, strategic network formation, and game-theoretic modeling (*Jackson*, 2010). In static random models, the network topology is decided by predefined rules with preexisting nodes, such as probability of any two nodes linkage is a constant p and n nodes are given from the beginning. One well-known model from this category is Poisson random graph model (*Erdos and Rényi*, 1960). Given its simplicity and early influence, it is often used as a benchmark model. While in growing random models, network formation is a dynamic process, and new nodes are added into the network in sequence. A famous model in this class is preferential attachment model (*Newman*, 2001b). Preferential attachment model overcomes Poisson random graph model's missing "fat tail" with increased probability of new node linking to highly connected existing nodes. One important dimension missing from these two groups of random models is that, in many setting, it is not just chance deciding the linkage, but choice plays a central role in linkage formation. In order to measure this additional dimension, a utility function is usually required. This utility function represents benefit of nodes from the network. These category of models are often referred as strategic network models. Pairwise stability model considers network formation as the state where every node achieves its marginal optimal connections under a certain utility function assumption (*Jackson and Watts*, 2001). Efficient network models, in comparison, is optimized for overall network benefits (*Newman*, 2001b). Different from efficient network models, game-theoretic models assume benefit conflicts between nodes

in the network and each node seeks its own benefit.

Food flow network has its own unique characteristics that are unnecessarily consistent with the well-studied networks by existing models. For instance, one prevailing property of these networks is "rich-gets-richer" (*Jeong et al.*, 2003). In a social network, it could be explained as an individual with bountiful friends would get more "friend-of-friend". However, in food flow network, these cyclical relationships would not makes sense (i.e. a food exporter sends food to one country, who then sends it to another country, before finally returning it to the origin country. Explanation of observations in food flow networks, has to be combined with its context. It has to consider environmental factors like food production and consumption, available transportation infrastructure, and spatial specialization of production, rather than adopting general network growth models. The literature specifically attempting to understand food flow networks has been rare. All these statistical network models are basically context-free. They are promised mainly to advance general network theory instead of towards explaining a specific network. A real-life network is supportive evidence to validate these models. However, our ideal model would be data-oriented, targeting specifically at the food flow networks we have. We need to explain the food flow network with its own characteristics, such as crop production, consumption, and transportation factors. Also, these models only concern with network topology, while food flow volume can not be represented in these models.

Compared to statistical network modeling, link prediction literature is usually context based (*Lü and Zhou*, 2011). According to (*Lü and Zhou*, 2011), link prediction methods can be broadly classified into three groups, similarity based, maximum likelihood based, and probabilistic. Some methods attempt to predict not only links, but also link intensities (i.e. "strength") (*Leroy et al.*, 2010). These methods tends to have high prediction accuracy, but usually lack the ability of explanation and don't consider preservation of overall network characteristics.

1.3 Intellectual merit

The primary goal of this dissertation is to understand the scientific principles of the food commodity flows. The overarching questions are:

- (1) What are the network properties of food flows within a country?
- (2) How do food flows vary with spatial scale?
- (3) How can we model food flows in locations without empirical information?

Based on understanding discovered through question (1) and (2), we create a data-driven model to quantitatively understand food flow networks, that will have custom interpretation towards food flow network, modeling both topology and strength, and preserve network characteristics to answer question (3). As an important practical application of this understanding and a validation of effectiveness of this model, county level food flow networks will be reconstructed, demonstrating this model preserves the network attributes observed in different spatial scales.

To address these research questions, this dissertation primarily uses network statistics. Different network attributes are studied across multiple scales to extract network structural properties. Once the empirical network patterns have been adequately characterized, this dissertation proposes a novel methodological framework to address question (3). A combination of mass balance, network constraints, machine learning, and linear programming is used to infer food flows in locations without data. The regression model underpinning this approach is based on the gravity model of trade (*Burger et al.*, 2009), which is an empirical relationship to explain the commodity flow between any two trading partners.

Network analysis of food flow network at global scale has been conducted in literature. However, such a analysis at a country scale was missing. In this dissertation, network analysis has been adopted to understand food flow network in USA, a major food producer in the work. After that, food flow network across multiple scales, including countries in the globe, sub-national locations in the USA, and households in villages, has been investigated. By comparison of food flow networks across multiple scales, the invariant and variant characteristics in these networks has been summarized. Based on the invariance properties, statistical modeling becomes feasible. There is little literature using context-based model that is both interpretable and predictive towards understanding food flow network topology and strength. The last section proposed stochastic processes that might be responsible for forming food flow networks, including both topology and flow strength. It also attempts a custom explanation to interpret these processes with context variables, such as food production, consumption, and other supply chain factors. It created a regression model of food flows between counties in the USA. This model can both produce network topology and flow strength while preserving observed network properties. To be more realistic, given this simulation result, it utilized empirical information in some areas as mass balance constraint to generate a food flow network with additional realism.

1.4 Research questions and objectives

1.4.1 Food flow in USA

Use network analysis to characterize food flow within USA and compare it against global food flow.

Research questions What are the network properties of food flows within USA? How clustered is the network? How assortative is the network? Are there hubs in the network? Is the food flow network at "social" network, a "technological" network or a mix? Does the food flow equally among the areas?

1.4.2 Scaling property of food flow networks

At village level, national level and international level, it studies all food flow patterns with network analysis. It finds the invariance and variance across scales. It tries to understand what are the processes generating this food flow system and propose stochastic process underneath that resulting in their characteristics. It attempts to interpret these processes within food flow context.

Research questions How do the network properties of food flow networks vary with scale? Are there consistent patterns across all scales? What stochastic processes will generate the topology and strength of the food flow network across scales?

1.4.3 Food flow between counties in United States

Based on assumptions of the underlying statistical processes generating the food flow network learned from last section, create a model to simulate food flow network at county level within the United States. Evaluate performance and verify that the network characteristics are preserved.

Research questions How can we model food flows in locations without empirical information? What are the drivers of the food flow network system? How are these drivers related to each other? How to build a model that is context-based, interpretable and network properties preserving for county level food flow network within the United States? How to simulate a network without knowing either network topology or link strength? How to interpret the model?

1.4.4 Research Contributions

1. This study presented a novel application of network analysis to domestic food flows within the USA. It revealed an unassortative network structure with high clustering

coefficients. These network properties indicated that the USA food flow network is highly social and well-mixed. However, a power law relationship between node betweenness centrality and node degree indicated potential network vulnerability to the disturbance of key nodes. This study showed that food flow network show similar equality metrics between USA and global.

2. This study presented an empirical analysis of food commodity flow networks across the full spectrum of spatial scales: global, national, and village. This study discovered consistent structural network properties of food flow networks across multiple scales, indicating that a universal mechanism may underpin food exchange systems. This study discussed the potential underlying stochastic processes responsible for these invariance network properties.

3. This study developed the Food Flow Model, a data-driven methodology to estimate spatially explicit food flows for subnational locations without flow data. This context-based model is both interpretable, and preserves the network properties observed across scales. This is the first model developed specifically for food flow network and combines both data and interoperability. Also it solved the topology problem and flow strength problem altogether. This model has important potential application in understanding food flow network and simulating flow flow network in areas where data is missing.

1.5 Dissertation structure

Chapter 2 Network analysis applied in food flow network in USA. Result is compared against global food network.

Chapter 3 Network analysis is conducted food flow networks across multiple spatial scales. Scale variant and invariant network properties are discussed.

Chapter 4 Take the statistical distributions in Chapter 3 one step further to develop models and simulate food flow network between counties in USA. It has shown preservation of network properties in the simulated network.

CHAPTER 2

A NETWORK ANALYSIS OF FOOD FLOWS WITHIN THE USA

2.1 Introduction

Food security is being placed under increased pressure due to economic development, population growth, and climate change (*Godfray et al.*, 2010; *Kastner et al.*, 2012). The world food system is increasingly globalized and inter-connected (*Ercsey-Ravasz et al.*, 2012), making it imperative to understand the consequences of this increasingly complex food system for a secure food supply. Trade flows are an essential component of the new, globalized food system and are increasingly important for global food availability (*Burgess and Donaldson*, 2010; *Porkka et al.*, 2013), with repercussions for carbon emissions (*Peters et al.*, 2011), nutrients (*Schipanski and Bennett*, 2012; *O'Bannon et al.*, 2013), water resources (*Konar et al.*, 2012), and povery (*Hertel et al.*, 2010). Thus, it is increasingly important to understand the structure of food trade. In this paper, we apply tools of network theory to domestic food flows within the USA, a country with global importance as a major agricultural producer and trade power (*Konar et al.*, 2011; *Dalin et al.*, 2012).

Network analysis has been increasingly used to understand complex systems. This recent interest in complex networks is largely due to the discovery of organizing principles in networks (*Newman et al.*, 2006; *Costa et al.*, 2007), such as community structure (*Watts and Strogatz*, 1998) and scale-free properties (*Barabási and Albert*, 1997). Additionally, network analysis has become increasingly popular due to its flexibility and ability to represent many real-world systems (*Barabasi*, 2002), including transporation systems (*Kalapala et al.*, 2006; *Masucci et al.*, 2009; *Kaluza et al.*, 2010), the world wide web (*Barabási and Albert*, 1997), international tourism (*Miguens and Mendes*, 2008), financial transactions (*Garlaschelli and Loffredo*, 2005; *Kyriakopoulos et al.*, 2009), and scientific collaborations (*Newman*, 2001a; *Barrat et al.*, 2004), among others. In this paper we present a novel application of network analysis to food flows within the USA.

Global trade has been studied for quite some time (*Tinbergen*, 1962), more recently using tools of network theory (*Serrano and Boguna*, 2003; *Garlaschelli and Loffredo*, 2005; *Bhattacharya et al.*, 2007; *Fagiolo et al.*, 2008; *Barigozzi et al.*, 2010). Recent work has begun to

focus on the network characterization of global food trade (*Konar et al.*, 2011; *Ercsey-Ravasz et al.*, 2012; *Shutters and Muneepeerakul*, 2012). This study advances reseach in this area in three main ways. First, this network analysis of domestic food flows occurs at a different scale to the studies of global trade networks in the literature. In this way, this study help us to understand the impact of scaling on network properties, which is an important question in the literature (*Serrano and Boguna*, 2003). Second, food flows within the USA occur without barriers to their movement (i.e. due to the Commerce Clause of the U.S. Constitution), thereby proxying a free trade setting. Studying the network properties of food flows within the USA can thus help us to understand the network properties that may occur under free trade situations. Third, flows of food within the USA serve as a null model for trade equity (i.e. since the USA has a homogeneous population, shared agricultural policy, and absence of trade barriers). In this way, studying the equity of food flows within the USA enables us to quantify how equitable we can expect global flows to be, which is an important focus of current research (*Hertel et al.*, 2010; *Seekell et al.*, 2011; *Konar and Caylor*, 2013; *Porkka et al.*, 2013).

2.2 Methods

2.2.1 Food flow data

We obtain data on the movement of food within the USA from the Commodity Flow Survey (CFS). The CFS was created through a partnership between the Census Bureau and the Bureau of Transportation Statistics. The CFS present information on the movement of commodities within the United States. It provides information on category, shipment value, weight, and mode of transporation for commodities originated from mining, manufacturing, wholesale, select retail and services establishment to their destinations. The CFS is conducted every five years as part of the Economic Census (*CFS*, 2013). However, bilateral data is only available for the year 2007 when this research is conducted, so we focus our analysis on this year. We select five categories of food commodities in CFS for our analysis. They are 'cereal grains', 'other agricultural products', 'animal feed and products of animal origin, nec', 'meat, fish, seafood, and their preparations', and 'other prepared foodstuffs and fats and oils'. For the USA, data is provided at both the state and 'CFS area' level. A CFS area is a geographic area that is drawn from a sub-set of Combined Statistical Areas (CSAs) and Metropolitan Statistical Areas (MSAs) as defined by the Office of Management and Budget. If a metropolitan area spans multiple states, then the CFS area is defined for each state part with significant transportation related activity. State parts of otherwise included metropolitan areas with little tranportation activity are included in the remainder

of that state. The CFS defines the 'Remainder of state' to represent those areas of a state not contained within a separately published metropolitan area. There are a total of 123 CFS areas for the 2007 database. They are listed in the Appendix A.1.

2.2.2 Network construction

We create bilateral matrices of food trade connections and volume flows within the USA. The CFS provides 15,512 data entries for food flow in value terms and 12,672 data entries for food flow in volume terms.

CFS network construction

We obtain commodity-specific weighted and directed matrices of food flows within the USA for 2007. The nodes of the network are the CFS areas of the USA and the links connecting nodes are weighted by the volume of food flow [tons] and directed by the direction of flow. The individual commodity-specific matrices are summed to obtain the aggregate food flow matrix. For the remainder of the paper, we refer to this network as 'aggregate'. Unless we specifically refer to the network of a certain commodity, we are referring to the aggregate food flow network. Most study in this paper is based on this network.

State network construction

By aggregating commodity flows from CFS areas within the same state, we can construct a network as each node represent a state. This network is utilized in the following sections where it is specified.

2.2.3 Network statistics

From the weighted and directed matrix of food flows (W), we calculate network statistics. Node degree is a fundamental network property that considers node connectivity. Specifically, node degree measures the total number of links of a node. This is an unweighted property, so we refer to the adjacency matrix (A). Since our network is directed, we consider node in- and out-degree, based on whether the import or export relationship is being considered, respectively. The node in-degree counts the number of links incoming to a node and is measured by $k_{in_i} = \sum_j a_{ji}$, while node out-degree counts the number of links emanating from a node and is measured as $k_{out_i} = \sum_j a_{ij}$, where a is an element of A (*Wasserman and Faust*, 1994).

To consider the weights assigned to links in our network, we quantify node strength. Node strength is the weighted corollary to node degree and measures the sum of the weights for nodal links. To take direction into account, we consider node in- and out-degree, as before. Now, node in-strength sums the value of links incoming to a node and is measured by $s_{in_i} = \sum_j w_{ji}$, while node out-strength sums the value of links emanating from a node and is

measured with $s_{out_i} = \sum_j w_{ij}$, where w is an element of W (*Wasserman and Faust*, 1994). Thus, node strength differentiates between connections with different values or intensities. Here, the volume of food trade [tons] provides the weight for our trade links.

Node degree and strength provide local measures of the importance of a node to the network. To better understand the importance of a node to the overall structure of the network, we consider average nearest neighbor degree, clustering, and betweenness centrality. Average nearest neighbor degree (*knn*) measures the affinity of a node to connect to high- or low-degree neighbors, or the network correlation structure (*Watts*, 1999; *Jackson*, 2008). When direction is taken into account, weighted values of *knn* can be measured with four directional pairs: in-in (ii), out-out (oo), in-out (io), and out-in (oi). The clustering coefficient (C) measures the degree to which nodes tend to cluster together or form closed triangles (*Watts*, 1999). With direction, there are eight possible combinations of C that fall into four categories (see (*Fagiolo*, 2007) for a complete description and representation): C_{in}, C_{out}, C_{cyc}, and C_{mid}. Our equations for *knn* and C are provided in the Appendix and follow (*Konar et al.*, 2011). Betweenness centrality (B) quantifies the importance of a node or link in terms of its importance to the overall network architecture. Here, we calculate node B which counts the fraction of shortest paths going through a given node, defined as $B = \sum_{i,j} \frac{\sigma(i,u,j)}{\sigma(i,j)}$, where $\sigma(i,u,j)$ is the number of shortest paths between nodes i and j that pass through node u, $\sigma(i,j)$ is the total number of shortest paths between i and j, and the sum is over all pairs i,j of nodes (*Costa et al.*, 2007). We normalize B by $(N-1)(N-2)/2$ to maintain $B \in [0,1]$ as suggested by (*Barthelemy*, 2004). Directed paths are used to calculate directed B and undirected paths for undirected B. B is an important measure of how important a node is for connecting other nodes in the network (*Jackson*, 2008). Finally, we conduct a triadic analysis of the USA food flow network. Triads are three-node directed sub-graphs. A small number of triad patterns are able to describe a wide variety of real-world networks (*Milo et al.*, 2004; *Shutters and Muneepeerakul*, 2012). Triad frequencies of empirical networks are compared to frequencies in a random network to arrive at a normalized z-score for each triad type (refer to Appendix). When the normalized z-score is plotted for all triad types, the triad significance profile (TSP) is obtained, which can be directly compared across networks.

2.2.4 Measures of equality

We calculate several measures for the equality of food flows across CFS areas in the USA. First, we calculate the Gini coefficient (G) (*Gini*, 1909; *Seekell et al.*, 2011). The Gini coefficient measures the inequality among values of a frequency distribution. $G \in$ [0,1], where 0 indicates perfect equality (i.e. all values are identical) and 1 indicates perfect inequality (i.e. one node has all of the value in the network). Next, the Lorenz asymmetry coefficient (S)

measures the assymetry of the Lorenz curve, which describe the inequality in the distribution of a quantity (*Damgaard and Weiner*, 2000; *Seekell et al.*, 2011). The Lorenz asymmetry coefficient is a useful corollary to the Gini coefficient. S values of 1 indicate that the Lorenz curve is symmetric, while values of $S > 1$ indicate a few nodes consuming more resources and $S < 1$ indicates inequality due to a large number of nodes with small food flows. Finally, we calculate the Hoover index (D), equivalent to the maximum vertical distance between the line of equality and the Lorenz curve. The Hoover index can be interpreted as the proportion of food trade by above-average nodes that would need to be redistributed to below-average nodes to achieve trade equality (*Hoover*, 1941; *Seekell et al.*, 2011).

2.3 Results

Summary statistics We present a map of the USA food flow network among states in Fig 2.1. This image illustrates the flows of food between states in the USA. The states are ranked according to their total food trade volume and plotted clockwise in descending order. The size of the outer bar indicates the total trade volume of each state as a percentage of total USA trade. Export volume is indicated with links emanating from the outer bar of the same color. Import volume is indicated with a white area separating the outer bar from links of a different color. The total trade volume of food represented by this graph is 829.3 x 10^6 tons. Note that the largest link in the USA food flow network is from Illinois to Louisiana. Midwestern states are shown as major exporters of food to the key ports in Louisiana, California, and Texas. In other words, the movement of food from the Midwest to ports in Louisiana, Texas, and California are key pathways for domestic food flows before international export. The dense web of connections illustrates how inter-connected trade between states is, which is what is expected, given the US Constitutional requirement for free trade between states. There are 123 nodes (i.e. CFS areas) and 8,396 links in the aggregate USA food flow network. Thus this is a dense network, since network density p is measured by $p = M/[N(N-1)]$, where M is the number of links and $N(N-1)$ is the number of possible links. Here, $p = 0.56$, compared with 0.33 for global food trade. In other words, the USA food flow network is more inter-connected than global food trade. Network summary data are provided by crop in Table A.6.

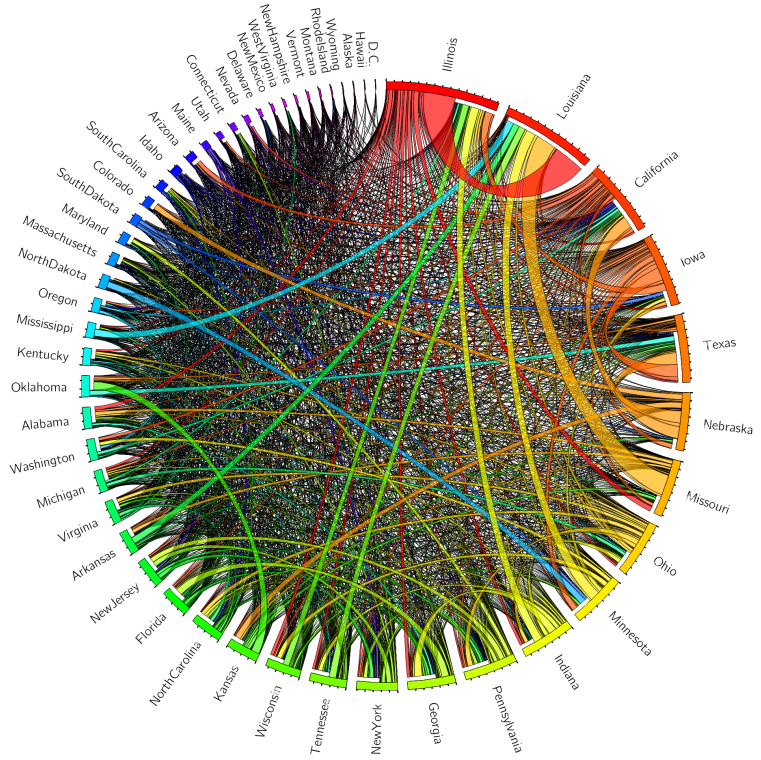

Figure 2.1: Network representation of food flows between the 50 USA states. The states are ranked according to the total trade volume and plotted clockwise in descending order. The size of the outer bar indicates the total food flow volume of each state. Export volume is indicated with links emanating from the outer bar of the same color. Import volume is indicated with a white area separating the outer bar from links of a different color. Note that the largest link in the USA food flow network is from Illinois to Louisiana. Midwestern states are shown as major exporters of food to the key ports in Louisiana, California, and Texas.

Degree and strength A highly skewed degree distribution is a common feature of many real-world networks. Power law degree distributions are a feature of some networks (*Barabási and Albert*, 1997), while deviations from power-laws (*Newman*, 2004), exponential degree distributions (*Guimera et al.*, 2006), and normal degree distributions (*Shutters and Muneepeerakul*, 2012) have also been shown. From Fig 2.2A,D it is clear that the USA food flow network exhibits a normal distribution, the hallmark of social networks amongst people (*Pennock et al.*, 2002). This differs from the scale-free character of the world trade web of all commodities (*Serrano and Boguna*, 2003). For global food trade specifically, the import degree distribution is also normal (*Shutters and Muneepeerakul*, 2012), similar to USA-only food flows, while the global export degree distribution exhibts a fatter tail than the USA food flow network. The average node degree in the USA food flow network is 68. The max-

imum in-degree is 86 and the minimum in-degree is 0. The out-degree ranges from 1 to 94. Los Angeles-Long Beach-Riverside' has the highest node in-degree, with 86 connections. The second and third most connected nodes in terms of in-degree are 'Chicago-Naperville-Michigan City' and 'Atlanta-Sandy Springs-Gainesville' with 83 and 81 connections, respectively. The 'Remainder of Wisconsin' has the highest out-degree of 94, while the second highest is 'Chicago-Naperville-Michigan City' and 'Iowa' with 88 connections. Refer to the SI for the top 10 connected nodes. 'Idaho' has the smallest in-degree, while 'Corpus Christi-Kingsville' has the smallest out-degree. The distribution of node strength for the USA food flow network is shown in Fig 2.2B,E. A Weibull distribution is fit to the data. The Weibull distribution indicates high heterogeneity in volumes of food movement around the USA, specifically in terms of export volumes. The equation for the Weibull distribution fit to s_{out} is $P(S_{out} > s_{out}) = e^{-\frac{s_{out}}{2.64}^{0.7}}$. The Weibull distribution provides the best fit to s_{in} ($P(S_{in} > s_{in} = e^{-\frac{s_{in}}{2.86}^{0.8}}$), although the left tail of the data diverges from the analytical distribution, indicating that more nodes of small import volume exist in the data than expected from the Weibull distribution. The mean node in-strength in the USA food flow network is 3.4 x 10^6 tons. The maximum node in-strength is 43.7 x 10^6 tons and the minimum is 0. The node out-strength ranges from 0.007 to 31.6 x 10^6 tons, with a mean value of 3.4 x 10^6 tons. 'New Orleans-Metairi-Bogalusa' exhibits the highest in-strength of 43.7 x 10^6 tons. The second and third highest in-strength are for the 'Remainder of Texas' and 'Los Angeles-Long Beach-Riverside', with values of 18.1 and 17.1 x 10^6 tons, respectively. The largest export volume is for 'Iowa' of 31 x 10^6 tons, followed by the 'Remainder of Illinois' and 'Remainder of Missouri', with volumes of 28.3 and 20.8 x 10^6 tons, respectively. Refer to Table A.7 for the list of the top 10 nodes in terms of strength.

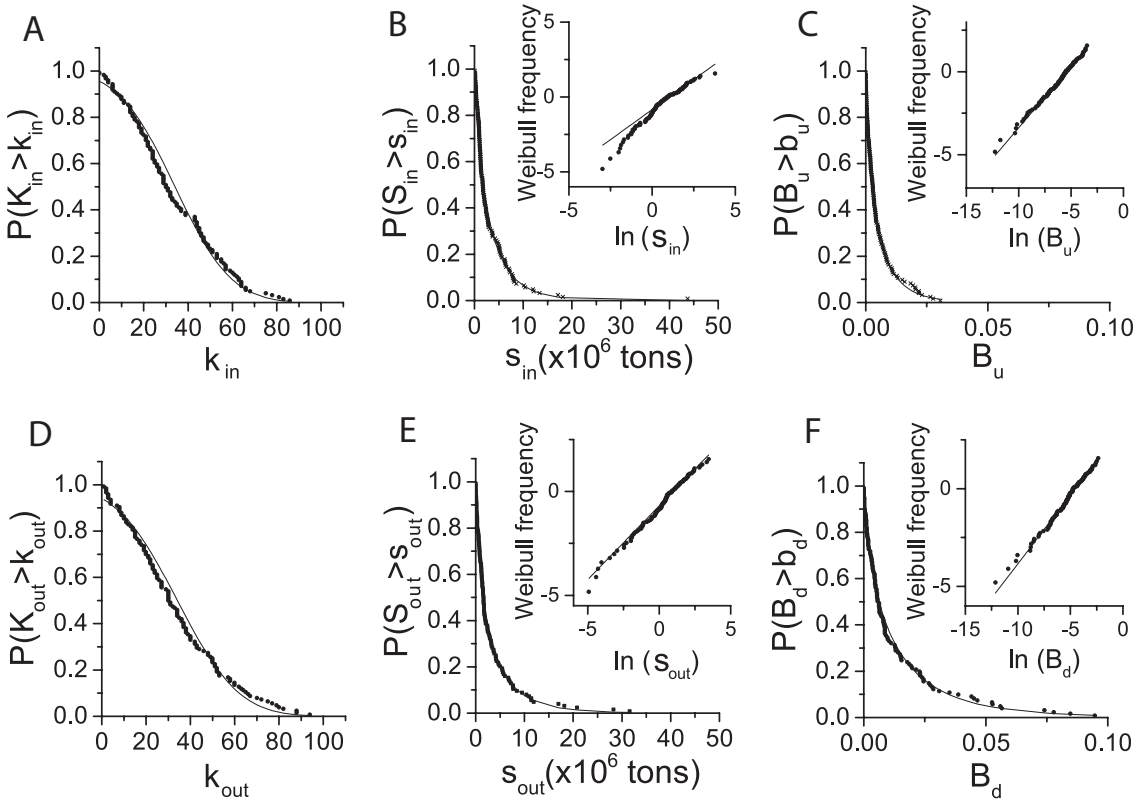

Figure 2.2: Node degree (k), strength (s), and betweenness centrality (B) distributions. In all plots the grey points indicate data and the black lines indicate analytical distributions. Import (Panel A) and export (Panel D) node degree distributions follow a normal distribution, reminiscent of the small world networks typical in social systems. The Weibull distribution fits export volume (Panel E; $P(S_{out} > s_{out}) = e^{-\frac{s_{out}}{2.64}^{0.7}}$), undirected B (Panel C; $P(B_u > b_u) = e^{-\frac{b_u}{0.004}^{0.7}}$), and directed B (Panel F; $P(B_d > b_d) = e^{-\frac{b_d}{0.01}^{0.7}}$), indicating higher heterogeneity than k with a 'fat' tail. Import volume (Panel B; $P(S_{in} > s_{in}) = e^{-\frac{s_{in}}{2.86}^{0.8}}$) is best fit by the Weibull distribution, but diverges from Weibull for low values of s_{in}, indicating that nodes with small numbers of import partners are more common in the data.

The relationship between node strength and node degree is shown in Fig 2.3, Note that the axes are in log-log and that the best fit to the data is linear, indicating a power law relationship between node strength and node degree, similar to global food trade (*Konar et al.*, 2011). This means that as a node increases its connectivity with other nodes, it is much more likely to trade larger volumes of food. The power law relationship between trade connections and volume is essentially independent of direction. However, the export trade relationship does display a slighter larger exponent, which differs from global trade (*Konar*

et al., 2011). The exponent for the power law relationship is very similar across commodities with the exception of import flows of cereal. Increasing import connections leads to much higher trade volumes of cereal, which is important to note since cereal plays such a dominant role in food security and global trade systems, particularly maize exports from the USA to the rest of the world (*Wu and Guclu*, 2013).

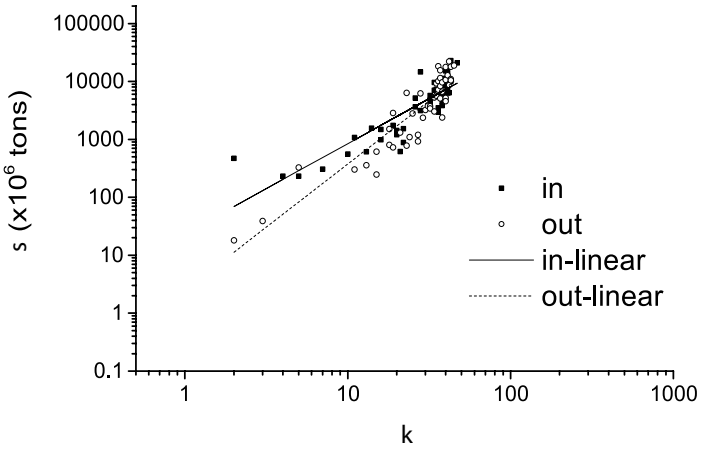

Figure 2.3: Relationship between node strength and node degree for food flows in the USA. The exponent for s^{in} vs $k^{in} = 1.335$ and the exponent for s^{out} vs $k^{out} = 1.555$.

Assortativity In the previous section we analyze node degree and strength, which are first-order network indicators. In other words, statistics on k and s only provide information on a node and its trade partners that are one step away, but do not contain information on the neighbors of that node or on the global network topology. In this section we investigate network assortativity, which is a second-order network indicator, since it includes information on nodes lying two steps away from the one under consideration (*Fagiolo et al.*, 2008). Network assortativity describes how similar connected nodes are in terms of some attribute. Here, we consider how similar the degrees of connected nodes are, i.e. assortative mixing by degree.

A common method for determining network assortativity is by plotting *knn* vs. degree. If this graph exhibits an increasing relationship, the network is referred to as 'assortative'. However, if a negative relationship is evident, then a 'dissassortative' network structure exists. Social networks tend to exhibit assortatvity, while technological and biological networks are most often characterized by disassortativity (*Newman*, 2002). Interestingly, economic networks exhibit features of both technological and social relationships (*Jackson*, 2008). For example, the venture capitalist network demonstrates positive degree correlations (*Mas et al.*, 2007), while negative degree correlations were shown for bank networks (*May et al.*, 2008)

and global trade (*Serrano and Boguna*, 2003). The network topology of domestic food flows (i.e. the unweighted conectivity structure) exhibits the structure of an 'unassortative' network, i.e. neither assortative nor disassortative. This differs from the global trade network, whose network topology exhibits disassortativity (*Serrano and Boguna*, 2003).

Network assortativity can also be quantified through the Pearson correlation coefficient (τ) (*Newman*, 2002). Here, we measure τ for degree between pairs of linked nodes, where values $\in$ (-1, 1). Values of $\tau = 1$ indicate perfectly assortative mixing, while values of τ = -1 indicate perfectly disassortative mixing (*Fricke et al.*, 2013). For an unassortative network $\tau = 0$. The connectivity structure of the USA food flow network exhibits τ values that are very close to 0, indicating largely unassortative mixing,particularly in comparison to a τ value for global trade roughly equivalent to -1 (*Fagiolo et al.*, 2008). Interestingly, knn_{ii} and knn_{oi} are both much more strongly disassortative than knn_{io} and knn_{oo}. This indicates that when you look at the neighbors of a given node, they tend to have lower import trade connections. Even though the domestic network does become assortative when food volumes are considered, the difference between the unweighted and weighted relationship is not significant In particular, the difference between the unweighted and weighted assortative structure is not as severe as it is for global trade (refer to Fig 6 in (*Konar et al.*, 2011)). For global trade, the unweighted *knn* structure is clearly disassortative, but becomes assortative when trade volumes are taken into account. This indicates that certain nodes hoard the majority of the resources amongst themselves at the global scale. This sharp difference in unweighted and weighted assortativity structure is not evident in USA food flows, although a difference is present. This indicates that the 'weighted rich club' (*Fagiolo et al.*, 2008) feature of global trade is not prominent in domestic food flows. Thus, domestic food flows exhibit assortative sorting, much like social networks (*Newman*, 2002), without evidence of a weighted rich club.

Clustering Clustering is a network property that describes the propensity of nodes in the network to form closed triangles. Clustering can be measured both locally and globally for networks. The local clustering measure evaluates the embeddedness of particular nodes. The global clustering measure indicates overall clustering within the network. We evaluate both local and global network clustering in this section. Clustering for global trade exhibits high heterogenity, in which there is a power law relationship between node clustering and degree (*Serrano and Boguna*, 2003). Here, the directed clustering and degree relationshipsare absent of the power law property in the global trade network. The relationship between C and k for USA food trade exhibits a much more homogeneous network, with high values of clustering across values of k. The addition of trade volumes to our calculation of local clustering does not change the slope of the graphs significantly, another indication that USA

food flows do no exhibit a weighted rich club (*Fagiolo et al.*, 2008).

Global C values are higher for USA food flows than for global food trade (refer to Table 8 of (*Konar et al.*, 2011)). This indicates that nodes within the USA are more inter-connected than nations participating in international trade (note that even values of C_{cyc} and C_{mid} are higher for USA food trade, which is a rare clustering pattern in global trade). However, the exception is for C_{in}^{W} and C_{out}^{W}. Values of C_{in}^{W} and C_{out}^{W} are 0.82 and 0.78 for USA trade and 0.94 and 0.73 for global trade. This shows the propensity for certain nations to dominate weighted trade at the global scale. Global C values are not significantly different in unweighted and weighted terms for the USA network, providing further evidence for the absence of a weighted rich club.

2.3.1 Betweenness centrality

Betweenness centrality (B) is the highest order network measure, since paths of any length that pass through a given node are considered (*Fagiolo et al.*, 2008). B is defined to be the number of shortest paths from all pairs of nodes in the network that pass through a given node of interest (refer to Methods). Thus, B is a measure of how important a node is to the entire network architecture. The distribution of node betweenness for the USA food flow network is provided in Fig 2.2C. B is best fit by a Weibull distribution, indicating a 'fat tail' representation of a few key nodes to the network. Directed B is fit by $P(B_d > b_d) = e^{-\frac{b_d}{0.01}^{0.7}}$ and undirected B is best fit by $P(B_u > b_u) = e^{-\frac{b_u}{0.004}^{0.7}}$. Thus, Directed B exhibits a fatter tail than does undirected B. In other words, when direction is taken into account, some nodes increase in their importance to the topology of the network. The node that exhibits the highest directed B is 'Los Angeles-Long Beach-Riverside', with a value of 0.095. The node with the second and third highest directed B values are 'Chicago-Naperville-Michigan City' and 'Remainder of Texas', with values of 0.085 and 0.078, respectively. Refer to Table A.8 for the top 10 nodes.

Node betweenness centrality vs node degree exhibits a power law relationship, where nodes with a high degree are much more likely to have a high betweeness value. This relationship is present for both the undirected and directed network. Networks with such a highly non-linear relationship between B and k suggest the presence of a network 'core', much like in the global trade system (*Ercsey-Ravasz et al.*, 2012), making the network particularly vulnerable to failure (*Motter and Lai*, 2002; *Thai and Pardalos*, 2012).

2.3.2 Triad analysis

A triad analysis of global food trade recently revealed a unique 'superfamily' when compared with existing networks (refer to (*Shutters and Muneepeerakul*, 2012) for comparisons

of TSPs of many real-word networks). This unique triad significance profile (TSP) revealed properties of both biological and human social networks. The TSP for global food trade indicates an abundance of triad types '9' and '10' (refer to Fig 2.4 for triad diagrams), which are indicators of biological networks, and abundance of triad type '13' and lack of triad type '6', the hallmark of human social systems (*Milo et al.*, 2004; *Shutters and Muneepeerakul*, 2012). Since global food trade exhibits topological characteristics of both biological and human networks, it is not surprising that the TSP of global food trade display a combination of these network types (*Shutters and Muneepeerakul*, 2012).

We present the triad analysis for domestic food flows in the USA in Fig 2.4. The overall TSP structure compares well with that for global trade of food. However, the USA food trade network exhibits even stronger signals of a human social network than does global food trade (note that triad type '13' is more prevalent and triad type '6' is less prevalent in Fig 2.4B). Thus, trade connections within the USA are highly social, more so than global trade patterns.

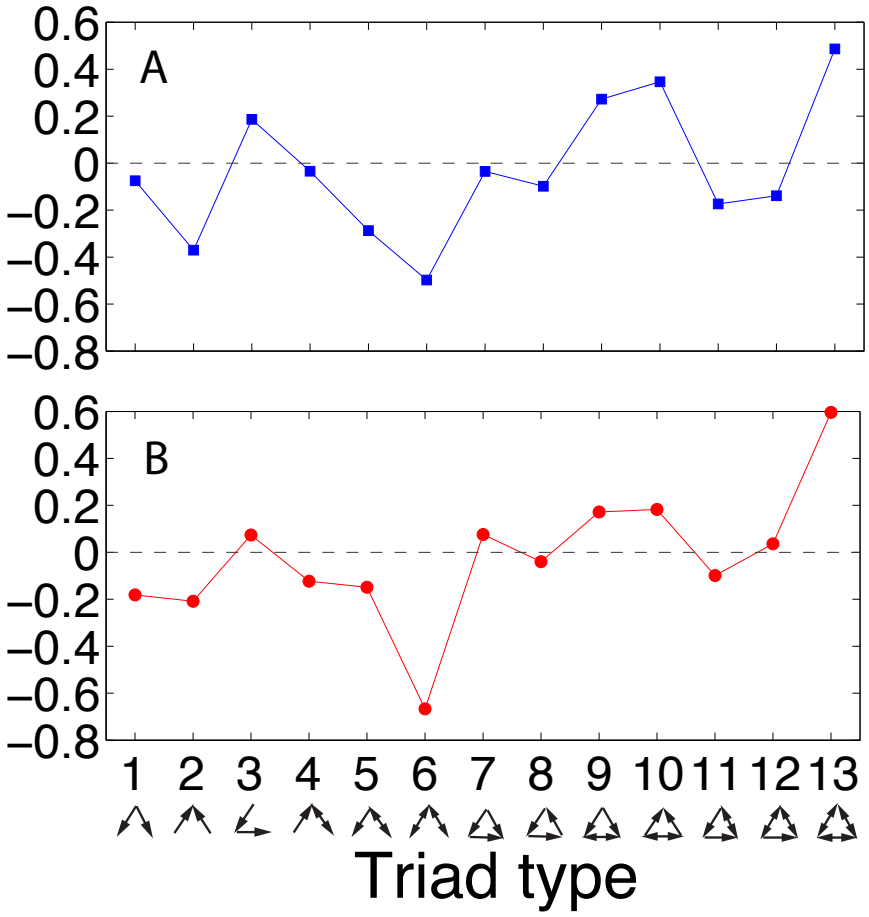

Figure 2.4: Triad significant profiles (TSPs) for food flows. (Panel A) TSP for global food flows, as presented in (*Shutters and Muneepeerakul*, 2012). (Panel B) TSP for food flows within the USA. The general structure of the triad significance profile is retained for domestic food flows. However, the USA food flow network exhibits more characteristics of a social network than does global food trade (i.e. anti-motif 6 and motif 13 are more pronounced in Panel B) .

Equality analysis Trade inequality is an important topic in the literature (*Seekell et al.*, 2011). In this section we compare statistics on global trade inequality with measures within the USA. We assume that food flows within the USA approximate the most equitable distribution that international trade may be expected to achieve. Thus, our analysis here serves as a benchmark for comparison for global trade equity. The Gini coefficient measures the inequality among values of a frequency distribution and has thus been often used to quantify trade equality. A Gini coefficient of 1 expresses the maximum inequality (i.e. one node has all of the wealth), while a value of 0 indicates perfect equality in the system (*Gini*, 1909). Additional statistics can be used to supplement the Gini coefficient in measuring distributional equality. The Lorenz curve asymmetry coefficient (S) describes the shape of

the distribution. An asymmetry coefficient of $S < 1$ describes a curve in which inequality is due to a large number nodes that import little. An asymmetry coefficient of $S > 1$ describes a curve in which inequality exists due to a few nodes dominating food trade. An asymmetry coefficient of $S = 1$ represents a symmetric curve (*Seekell et al.*, 2011). The Hoover index (D) quantifies the proportion of the value that would have to be redistributed to achieve perfect equality. If the entirety of food trade volumes would need to be redistributed to achieve perfectly equitable trade then we would obtain $D = 1$ (i.e. 100%). However, if perfectly equitable trade already exists, then we would not need to redistribute trade, so we would obtain a Hoover index = 0.

We perform an equality analysis which serves as a benchmark for global food trade, where the Gini coefficient = 0.579, Lorenz asymmetry coefficient = 0.966, and Hoover index = 0.442. Table 2.1 presents the Gini coefficient for food flows across the USA. These findings shed insight into trade network scaling and proxy free trade and equitable network architectures. The Lorenz curve asymmetry coefficient is essentially symmetric for domestic food flows in the USA, since $S = 0.97$ (refer to Table 2.1). D for USA food trade is 0.442, indicating that some trade would need to be redistributed to achieve perfect equality, although not as much as at the global level, in which half would need to be redistributed (*Seekell et al.*, 2011). Global food trade equality (measured in (*Seekell et al.*, 2011)) has been estimated with $G = 0.626$, $S = 0.70$, and $D = 0.5$. We think it is unlikely that international trade will be able to surpass the equality measures of USA food flows presented in this paper (where $G = 0.579$, $S = 0.966$, and $D = 0.442$). Thus, our analysis of domestic food flows raises the question of whether perfect equality is possible or even desirable within a trade system.

Table 2.1: Gini coefficient (G), Lorenz asymmetry coefficient (S), and Hoover index (D) for USA food flow networks. All other acronyms follow those in Table A.6.

	G	S	D
Aggregate	0.5788	0.9661	0.442
Cereal	0.9083	0.9578	0.7708
OthAg	0.6569	0.936	0.4848
Animal	0.6653	0.9552	0.5083
Meat	0.5844	0.8849	0.4415
Other	0.5379	0.9414	0.4022

Thus, the world food system has become increasingly complex and inter-connected, particularly due to food trade. The international trade of food commodities has been previously

studied using the tools of network analysis (*Konar et al.*, 2011; *Ercsey-Ravasz et al.*, 2012; *Shutters and Muneepeerakul*, 2012). In this chapter, we presented a novel application of network theory to food flows within a single country: the USA, a key nation in the global food network, since it is a major agricultural producer, consumer, and trade power. This analysis provides a useful benchmark for network properties across scales of trade, within a free trade setting, and for a relatively equitable case study. As expected, the USA food flow network is more equitable than global food trade. However, even food flows within the USA are not perfectly equitable and present a potential bound for how equitable global food trade can realistically be expected to be.

CHAPTER 3

SCALING PROPERTIES OF FOOD FLOW NETWORKS

3.1 Introduction

We live in an increasingly global society, in which food commodity transfers enable production and consumption activities to be separated in space via complex supply chains (*Seto et al.*, 2012; *Liu et al.*, 2013). We refer to the movement of food commodities from one location to another as 'food flows', reserving the term 'food trade' for the exchange of food commodities between locations. Recently, much research has evaluated food trade (*Ercsey-Ravasz et al.*, 2012; *Shutters and Muneepeerakul*, 2012; *Porkka et al.*, 2013; *Puma et al.*, 2015), particularly the resources embodied in food trade, such as water, carbon, and nutrients (*Peters et al.*, 2011; *Hoekstra and Mekonnen*, 2012; *MacDonald et al.*, 2012). However, we know relatively little about food flows at smaller spatial scales, such as within nations or cities. This is largely due to a lack of available data on food flows at smaller spatial scales. Research is needed to understand food flows across spatial scales in order to uncover the scientific principles behind food flows. To this end, we present an empirical analysis of food flows across the full spectrum of spatial scales: global, nation, and village.

Despite the importance of food flows at local to global scales, we do not understand how they are similar or different depending on the scale of analysis. Evaluating similarities and differences in food flows across spatial scales may yield important insights into their underlying structure and function. Scaling analyses have yielded insight into the underlying observed patterns and processes in ecology (*Levin*, 1992), biology (*West et al.*, 1997), hydrology (*Blöschl and Sivapalan*, 1995), and urban metabolism (*Bettencourt et al.*, 2007), among others. Within many local communities, it is common practice for some households to share food and other goods with those households that have experienced a negative shock event, such as a death in the family, outbreak of pests, or drought event (*Jackson et al.*, 2012; *Baggio et al.*, 2016). These household resolution exchanges of food within a village represent the smallest spatial scale of social food exchanges. At the global scale, nations trade food commodities according to their comparative advantage, resource endowments, food production and trade policies, and international politics. Food flow networks are driven by human

behavior across spatial scales, so the mechanisms leading to food flows relate to the food production, consumption, and trade decisions of people. In this way, the fundamental human behavior driving food flows (e.g. cooperation, kinship, risk sharing, economic welfare maximization, etc.) may leave its signature on the patterns of food flows across spatial scales.

Recent work has made significant strides in empirically describing (*Ercsey-Ravasz et al.*, 2012; *D'Odorico et al.*, 2014; *Lin et al.*, In Review) and modeling food flows (*Smith et al.*, 2017). Future food flow modeling would benefit from enhanced understanding of the empirical properties of food flows across spatial scales. In this paper, we empirically characterize food exchange networks across three dramatically different spatial scales: global, nation, and village. For the village scale, we obtain data on household resolution donations of food and non-food commodities within three Alaskan villages (*Baggio et al.*, 2016). Households exchange food items and other necessities with their neighbors and other households in their community in response to the heterogeneous distribution of availability and need, driven by a sense of kinship and reciprocity (*Nolin*, 2010; *Baggio et al.*, 2016). For the national scale, we obtain data on commodity flows between Commodity Flow Survey zones of the United States (*CFS*, 2013). The heterogeneous distribution of production and consumption at the national scale are additionally constrained by critical and inter-dependent infrastructure, such as the national transportation system (*Rinaldi et al.*, 2001; *Xu et al.*, 2012). For the global scale, we obtain data on international commodity trade from the United Nations Commodity Trade Statistics Database (COMTRADE) (*COMTRADE*, 2016). Heterogeneity in production and consumption and infrastructure constraints are again important, along with the additional forces of trade policies and global commodity markets (*Reimer and Li*, 2010). Network statistics provide a coherent framework for comparing complex systems across a range of scales (*Sayles and Baggio*, 2017). Examining how food flow networks change with the scale of description is essential in order to elucidate mechanisms underlying observed patterns, as well as for simplification, aggregation, and scaling (i.e., the relationship of variables with some measure of size or scale). Additionally, understanding network characteristics enables us to gain insight into the potential susceptibility of these interconnected commodity flow systems to shocks (*Puma et al.*, 2015). The network structures of food flow systems will provide a signature of their vulnerability and resiliency to disturbance (*Ercsey-Ravasz et al.*, 2012; *Puma et al.*, 2015), with important implications for embodied energy and water resources (*Melissa et al.*, 2017).

Network properties of food and embodied resource transfers have been investigated at the global (*Ercsey-Ravasz et al.*, 2012; *Konar et al.*, 2011; *Dalin et al.*, 2012), national (*Lin et al.*, In Review; *Dang et al.*, 2015), and urban (*Rushforth and Ruddell*, 2016; *Chini et al.*,

2017) scales. However, the network properties of the commodity flows that underpin virtual resource transfers have not been compared consistently across spatial domains. Similarly, comparisons between food and non-food flows, flow directionality (i.e. origin, destination, undirected), and unit of measurement (i.e. mass, value) have not been consistently evaluated. Here, we quantify food commodity flow network properties across the full spectrum of spatial scales. Importantly, we compare food flows networks with non-food commodity flows, by flow direction, and by measurement unit. The primary question we address is: How do food flow networks vary with spatial scale? We also address the following questions: How are food and non-food flow networks different? How does flow direction impact network properties? How does the unit of measurement impact network properties? We present our methods in Section 3.2. Our results are presented in Section 3.3. We conclude in Section 3.4.

3.2 Methods

Here, we describe the methods used in this paper. First, we detail the data sources of food flows at each spatial scale. Second, we explain the network statistics and distributions that we used to quantify these food flows.

3.2.1 Commodity flow data across scales

We obtain empirical information on commodity flows at three spatial scales: 'global', 'national', and 'village'. 'Global' data refers to international commodity trade between 240 countries for the year 2009. International trade data comes from COMTRADE (*COMTRADE*, 2016) and is mapped in Fig 3.1A. 'National' commodity flow data is for the United States and is obtained from the Commodity Flow Survey (CFS) for the year 2007 (*CFS*, 2013). The CFS dataset breaks the United States into 132 CFS Areas. A map of commodity flows within the United States is provided in Fig 3.1B. 'Village' scale data on commodity flows are available for all households for three Alaskan villages: Wainwright, Kaktovic, and Venetie (locations shown in Fig 3.1C). Data on village scale commodity exchanges are available for the years 2009 and 2010 (*Baggio et al.*, 2016) and are mapped in Fig 3.1C. Importantly, each dataset provides information on bilateral transfers between all nodes within the spatial domain, eliminating selection bias.

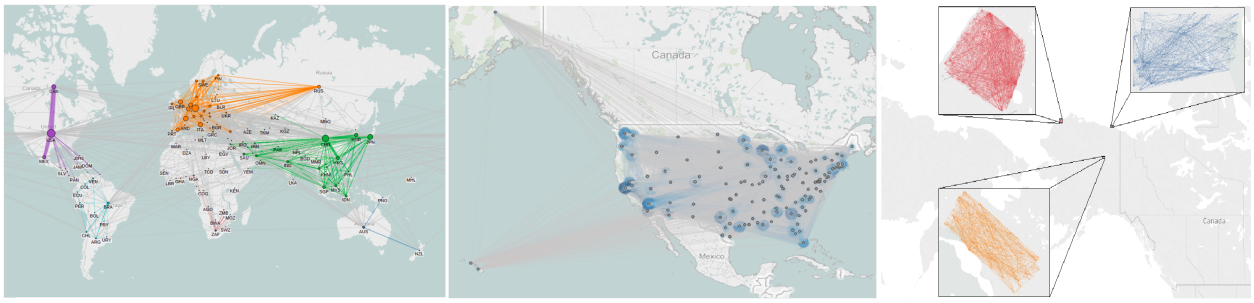

Figure 3.1: Maps of food flow networks for all spatial domains. (A) Map of international food trade between countries. (B) Map of food flows between Commodity Flow Survey areas in the United States. (C) Map of food exchanges between households in three Alaskan villages: (i) Wainwright, (ii) Kaktovic, and (iii) Venetie. Note that network illustration for villages is for visualization purposes only, as geographic locations of households are not provided. In (A) and (B) bubbles are scaled according to the total mass flux of food by node.

The average area of a country in the global trade system is 5 x 10^{11} m^2 (*CIA*, 2017). The area of the United States is 9,147,420 km^2 (*CIA*, 2017), and there are 132 CFS Areas. So, we estimate that the average area of a CFS Area in the United States is 7 x 10^{10} m^2. The average size of an American house is 222 m^2 (*Census*, 2017), which we assume is similar to the size of households in the three Alaskan villages. So, node size in the three systems varies across roughly 10 orders of magnitude.

Commodity flow data can be broken down into food and non-food commodities, enabling us to distinguish the unique aspects of food commodity flows across scales. At the global scale, Harmonized System (HS) codes 1 to 24 are food commodities, while non-food items are HS codes greater than 24. For the national scale, the Standard Classification of Transported Goods (SCTG) codes 2, 3, 4, 5, and 7 are food commodities and non-food commodities are all other SCTG codes. Village scale data on commodity donations are all food, with the exception of reports of equipment, cash, gas, ammunition, and donations of labor (*Baggio et al.*, 2016). Global and national commodity flows are provided in both mass [kg] and value [$], while village flows are only available in units of mass. We weight commodity flows in these primary units, as they are most likely to be linked to the underlying heterogeneity and driving mechanisms.

3.2.2 Network analysis of commodity exchanges

For each commodity flow network, the nodes are the units exchanging commodities and the links are the weighted, directed bilateral commodity flows. The adjacency matrix (A) is a binary matrix in which each element ($a_{i,j}$) is equal to 0 when no connection exists between nodes i and j and equal to 1 when there is a connection. Analogously, the weighted matrix

(W) contains elements $(w_{i,j})$ that provide the weighted link-level flows between nodes i and j. Network density refers to the number of links that exist in the network as a fraction of the total potential number of links. Density is a global network property and is measured by $p = M/[N(N-1)]$, where M is the number of links and $N(N-1)$ is the number of possible links (*Costa et al.*, 2007).

First order network properties examine the attributes of individual nodes in the network. Node degree (k) is a fundamental network property that measures the connectivity of each node, defined $k_i = \sum_j a_{i,j}$. So, k measures the number of commodity exchange partners of each node. Node strength (s) takes weight into account by summing the weights assigned to each node's links, $s_i = \sum_j w_{i,j}$ (*Costa et al.*, 2007).

Higher order network properties examine attributes of the neighborhood of nodes in the network. Node clustering (c) describes the propensity of nodes in the network to form closed triangles (*Watts*, 1999). This is a classic measure of the 'cliquishness' of a social network. Node betweenness centrality (B) is calculated as $B = \sum_{i,j} \frac{\sigma(i,u,j)}{\sigma(i,j)}$, where $\sigma(i,u,j)$ is the number of shortest paths between nodes i and j that pass through node u, $\sigma(i,j)$ is the total number of shortest paths between i and j, summed over all pairs i, j of nodes (*Costa et al.*, 2007). B is normalized by $1/(N-1)(N-2)$ such that $\in [0,1]$ (*Barthelemy*, 2004). Directed paths are used to calculate directed B and undirected paths for undirected B. B measures the importance of a node to the overall network structure.

3.3 Results and discussion

Here, we provide the results of our empirical analysis of both food and non-food flows at global, national, and village spatial scales. First, we map and determine the global properties of flow networks. Second, we determine the statistical distributions that best fit the networks across spatial scales. Third, we characterize the parameters of the statistical distributions across all scales.

3.3.1 Summary statistics

Fig 3.1 provides a map of food flows for each spatial domain of analysis. Panel A maps international food trade between countries. Note that world regions are color coded so that regional food trade can be more clearly observed. Panel B maps sub-national food flows within the United States. Food flows between the 132 CFS areas are depicted. Bubbles in Panel A and B are scaled by the total mass flux of food for each node. Panel C illustrates food flows between households in three Alaskan villages: (i) Wainwright, (ii) Kaktovic, and (iii) Venetie. However, geographical information is not available for households at the village spatial scale. So, maps of village food networks are provided for illustration purposes only

and are not geographically accurate.

Table 3.1 provides summary statistics for the three commodity flow networks. There are 240 nodes (i.e. countries) in the global food trade network, 123 nodes (i.e. CFS areas) in the national food flow network, and 163 nodes (i.e. households) in the village food flow network. Note that the 'village' provided in Table 3.1 is Kaktovic for comparison purposes. Kaktovice is representative of the other two Alaskan villages.

Table 3.1: Summary statistics for commodity flow networks. Statistics are presented for undirected food and non-food networks across spatial scales.

	Global	National	Village
Food			
# Nodes	240	123	163
# Links	13,438	3,002	628
Density	0.47	0.40	0.05
Mass [kg]	1.67×10^{12}	0.41×10^{9}	68,117
Value [$]	1.07×10^{12}	0.38×10^{12}	
Non-food			
# Nodes	240	123	163
# Links	17,160	5,824	384
Density	0.60	0.66	0.03
Mass [kg]	9.28×10^{12}	2.62×10^{9}	66,789
Value [$]	12.30×10^{12}	3.41×10^{12}	

The network properties of all Alaskan villages are provided in Table 3.2. Note that the number of nodes is constant across food and non-food flow networks. This indicates that all nodes trade both food and non-food. However, the number of links varies between commodity classes. Global and national domains have more non-food links, while the village domain has more food links.

Table 3.2: Summary statistics for village commodity flow systems. Statistics are presented for undirected food and non-food networks.

	Venetie	Wainwright	Kaktovi
Food			
# Nodes	205	217	163
# Links	560	1,063	628
Density	0.03	0.05	0.05
Mass [kg]	21,947	118,498	68,117
Non-food			
# Nodes	205	217	163
# Links	277	570	384
Density	0.01	0.02	0.03
Mass [kg]	13,862	190,425	66,789

Network density decreases with spatial scale for food commodities. This means that the fraction of realized to potential links declines as the spatial scale decreases. This relationship hints at scale dependence in food flow networks. Yet, density does not follow this clear pattern for non-food commodities, in which the density of the national scale is actually greater than it is at the global scale. Density of the village scale is dramatically lower than it is for the global and national scale. This is true for both food and non-food flows.

The mass and value of global trade is on the same order of magnitude for both food and non-food. This is quite surprising given the relatively low value of food commodities. Interestingly, national commodity flows in the United States are higher in value than they are mass for both food and non-food commodities. In fact, value flows within the United States are comparable to the entirety of the global trade system. This indicates that roughly the same value of commodities flows within a country as across all national borders, highlighting the importance of considering sub-national commodity fluxes.

3.3.2 Network distributions and parameters

Network connectivity

Fig 3.2A, D, and G present the degree distributions of undirected food flows across spatial scales. We fit a generalized exponential distribution to the node degree (k) distribution at each scale. The generalized exponential probability density function is given by:

$$f_i(x, \lambda_i, \lambda_{12}, s_i) = [\lambda_i + \lambda_{12}(1 - e^{-s_i x})]exp[-\lambda_i x - \lambda_{12} x + \frac{\lambda_{12}}{s_i}(1 - e^{-s_i x})] \quad (3.1)$$

This is referred to as the 'standardized' form of the generalized exponential probability den-

sity function. Location and scale parameters can also be used to shift and/or scale the distribution. Here, $f(x, \lambda_i, \lambda_{12}, s_i, location, scale)$ is equivalent to $f(x - location, \lambda_i, \lambda_{12}, s_i)/scale$ with $y = \frac{(x-location)}{scale}$ (*Ryu*, 1993; *Balakrishnan et al.*, 1998).

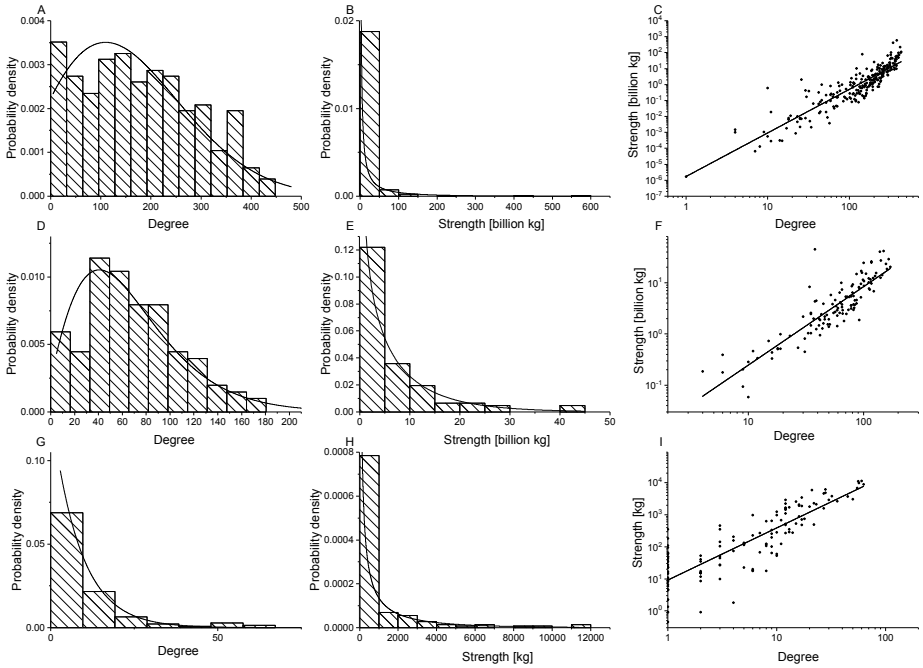

Figure 3.2: Network properties for undirected food flow networks. Global scale is shown in the top row (Panels A, B, C), national scale is shown in the middle row (Panels D, E, F), and village scale is shown in the bottom row (Panels G, H, I). Node degree distributions with generalized exponential distributions fit to the data are shown in the first column (Panel A, D, G), node strength [kg] distributions with gamma distributions fit to the data are shown in the second column (Panels B, E, H), and power law relationships for node strength versus degree are shown in the third column (Panels C, F, I).

Fig 3.2 A, D, and G illustrate that undirected food flow node degree distributions are well fit by a generalized exponential distribution across all spatial scales. However, Fig 3.3A, D, and G indicate that non-food flows are not well fit by the generalized exponential distribution. In particular, the right tail of the histogram for global and national connectivity exhibit higher values than can be captured by the generalized exponential distribution. This indicates that food and non-food commodity flows exhibit different network structures, likely due to differences in the underlying driving mechanisms. The generalized exponential distribution parameters for undirected food and non-food networks are provided by spatial scale in Table 3.3.

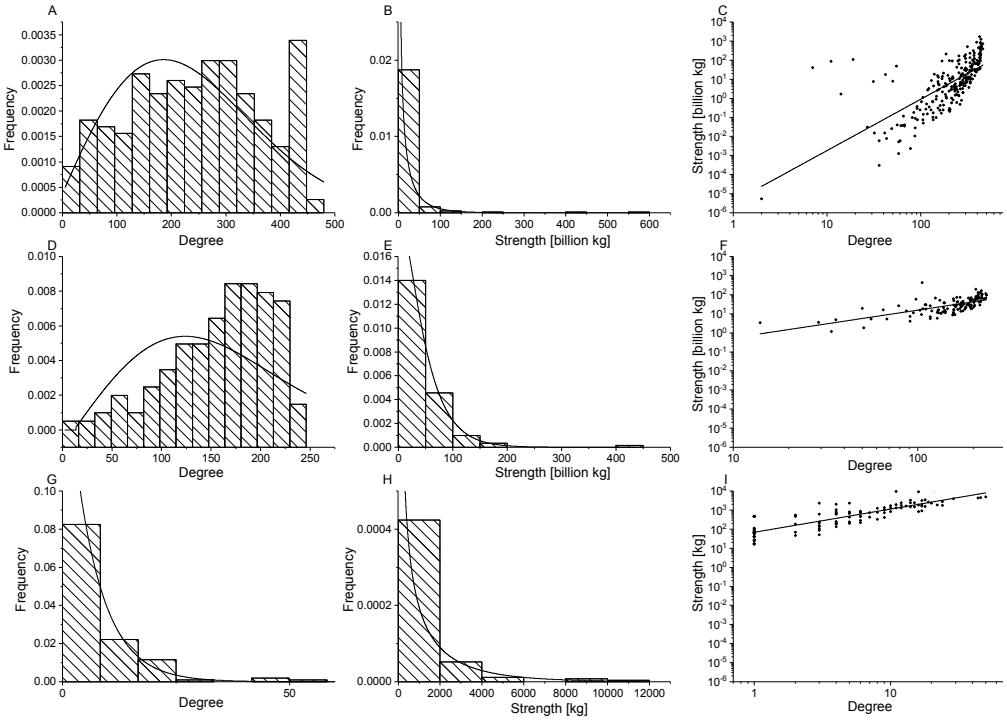

Figure 3.3: Network properties for undirected non-food flow networks. Global scale is shown in the top row (Panels A, B, C), national scale is shown in the middle row (Panels D, E, F), and village scale is shown in the bottom row (Panels G, H, I). Node degree distributions with generalized exponential distributions fit to the data are shown in the first column (Panel A, D, G), node strength [kg] distributions with gamma distributions fit to the data are shown in the second column (Panels B, E, H), and power law relationships for node strength versus degree are shown in the third column (Panels C, F, I).

Table 3.3: Parameters for food and non-food commodity flow networks. Parameters are presented for undirected food and non-food networks across spatial scales. Note that a generalized exponential distribution is fit to node degree, Gamma distribution is fit to node strength, and a power law is fit to node degree versus strength.

	Global	National	Village
FOOD			
Generalized Exponential Distribution			
λ_i	$8.45e-003$	$8.31e-003$	$2.78e+00$
λ_{12}	$7.80e-001$	$1.61e+00$	$2.25e-011$
s_i	$5.68e-004$	$2.35e-003$	$5.30e-001$
$location$	0	0	0
$scale$	$3.94e+00$	$3.69e+00$	$2.66e+001$
$KL-divergence$	0.54	0.39	1.22
Gamma Distribution			
α	$2.62e-01$	$8.28e-01$	$2.13e-01$
θ	$5.33e+01$	$8.15e+00$	$4.44e+03$
$KL-divergence$	1.17e-01	4.37e-02	1.72e-01
Power Law Fit			
a	-5.77 ± 0.18	-2.14 ± 0.13	0.96 ± 0.07
b	2.73 ± 0.09	1.53 ± 0.07	1.63 ± 0.09
R^2	0.81	0.78	0.72
NON-FOOD			
Generalized Exponential Distribution			
λ_i	$3.61e-04$	$2.89e-06$	$2.25e+00$
λ_{12}	$9.39e-02$	$4.28e-01$	$1.01e-11$
s_i	$5.88e-03$	$3.16e-03$	$4.49e-01$
$location$	0	0	0
$scale$	$4.30e+00$	$4.33e+00$	$1.64e+01$
$KL-divergence$	0.33	0.23	1.16
Gamma Distribution			
α	$2.70e-01$	$1.30e+00$	$3.60e-01$
θ	$2.86e+02$	$3.26e+01$	$2.97e+03$
$KL-divergence$	1.17e-01	9.24e-03	6.13e-02
Power Law Fit			
a	-5.43 ± 0.43	-1.72 ± 0.30	1.84 ± 0.05
b	2.69 ± 0.18	1.45 ± 0.14	1.22 ± 0.07
R^2	0.47	0.48	0.73

The generalized exponential distribution also provides a good fit to the connectivity of food flow networks with direction and value weights. When direction is taken into account, food flows are still well fit by the generalized exponential distribution. To see this, refer to Fig A.4A, D, and G and Fig 3.4 A, D, and G. This indicates that food flow networks can be modeled with the same distribution with or without consideration of flow directionality.

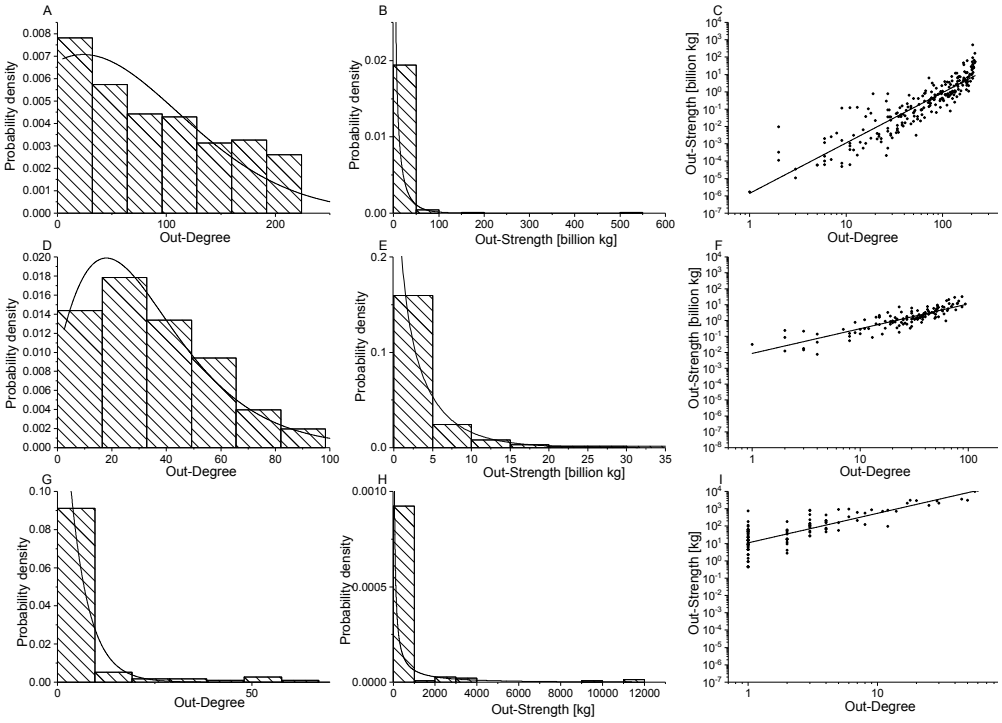

Figure 3.4: Network properties for export food flow networks. Global scale is shown in the top row (Panels A, B, C), national scale is shown in the middle row (Panels D, E, F), and village scale is shown in the bottom row (Panels G, H, I). Node out-degree distributions with generalized exponential distributions fit to the data are shown in the first column (Panel A, D, G), node out-strength [kg] distributions with gamma distributions fit to the data are shown in the second column (Panels B, E, H), and power law relationships for node out-strength versus out-degree are shown in the third column (Panels C, F, I).

Similarly, Fig 3.7A, D, and G illustrate that the generalized exponential distribution fits the connectivity structure of food flows well when value [$] weights are assigned rather than mass fluxes [kg]. This is what we would expect, since link weights are not considered in the connectivity structure, but it is good to empirically determine this. Additionally, all three Alaskan villages exhibit the same network properties (refer to Fig 3.6). However, Fig 3.5A illustrates that mean node connectivity decreases with spatial scale, despite the fact that the statistical distribution remains the same.

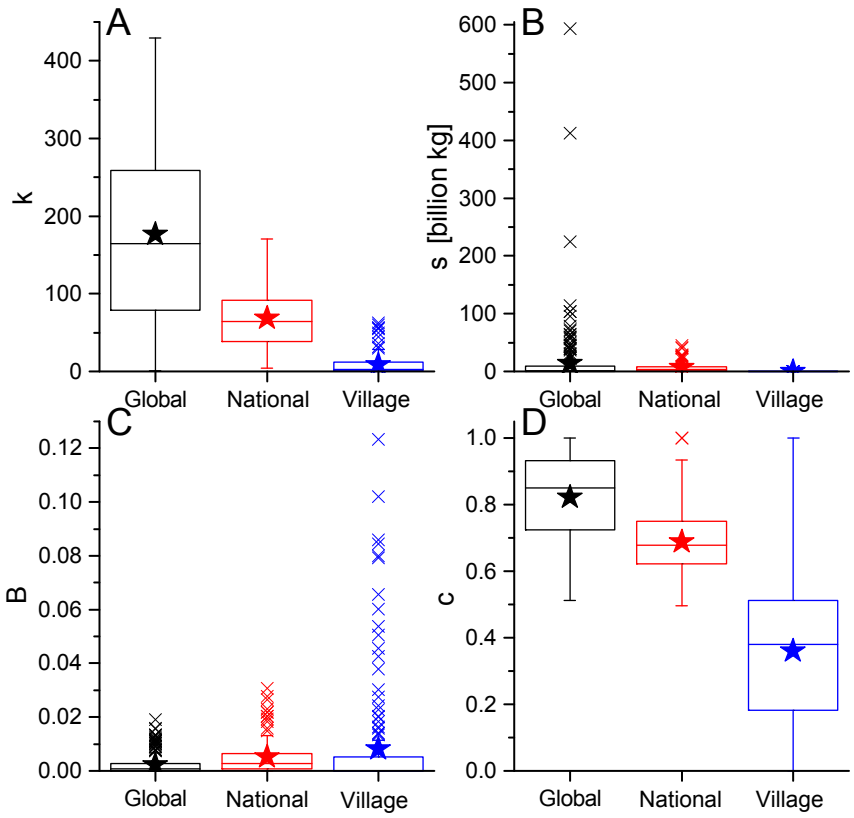

Figure 3.5: Network distributions for undirected food flows by spatial scale. Node degree (Panel A), strength (Panel B), betweenness centrality (Panel C), and clustering (Panel D) are shown for global, national, and village spatial scales. Box-whisker plots present the median (box line), interquartile range (box), mean (star), and outliers ("x").

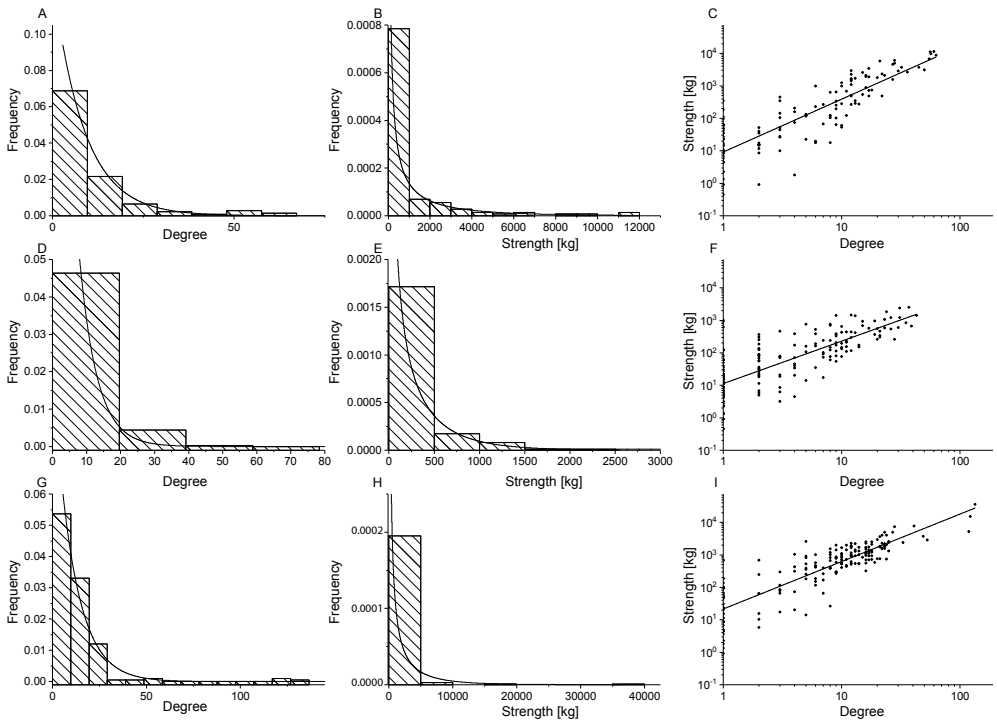

Figure 3.6: Network properties for undirected food flow networks of all villages. Kaktovic is shown in the top row (Panels A, B, C), Venetie is shown in the middle row (Panels D, E, F), and Wainwright is shown in the bottom row (Panels G, H, I). Node degree distributions with generalized exponential distributions fit to the data are shown in the first column (Panel A, D, G), node strength distributions with gamma distributions fit to the data are shown in the second column (Panels B, E, H), and power law relationships for node strength versus degree are shown in the third column (Panels C, F, I).

Like the Gamma and Weibull distributions, the generalized exponential distribution is an extension of the exponential distribution: if the shape parameter equals 1 then all three distributions coincide with the two-parameter exponential distribution (*Gupta and Kundu*, 1999). The generalized exponential distribution is not memoryless, such that shocks accumulate in time (*Ryu*, 1993). This makes sense, as disruptions to trade would likely be taken into account by trading partners going forward. This memoryless feature of the generalized exponential distribution leads to an increased failure rate when exposed to external shocks. A desirable feature to using the generalized exponential distribution to fit node connectivity is that it enables the shock arrival rates to be separately identified from their impacts (*Ryu*, 1993). This feature will likely prove useful in future research that aims to examine the impact of shocks to the global food trade system.

Network flows

The distributions of mass transfers for undirected food flow networks are provided in Fig 3.2B, E, and H. We fit a Gamma distribution to the node strength (s) distribution at each scale. The Gamma probability distribution function is:

$$\frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\theta}} \tag{3.2}$$

where Γ is the Gamma function, α is the shape parameter, and θ is the scale parameter of the Gamma distribution. Fig 3.2B, E, and H show that node mass flux, or strength (s), distributions for undirected food flow networks are fit well by a Gamma distribution across all spatial scales. This highlights that these networks are much more heterogeneous in terms of their mass fluxes than connectivity structure.

The Gamma(shape, success rate) distribution is generated from a Poisson process. Conceptually, the gamma distribution can be explained as commodity shipments to meet a "shape" amount of need. The chance that a unit of transported commodity will be successful and meet one unit of need is the "success" rate. The Gamma distribution parameters (α, θ) are provided for undirected food and non-food commodities in Table 3.3. Graphs of the Gamma distribution fit to non-food flow networks are provided in Fig 3.3. Fig A.4 and Fig 3.4 show the Gamma distribution fit to directed food flow networks. Node strength distributions in value [$] weights are provided in Fig 3.7 for the global and national spatial scales. These figures all indicate that the Gamma distribution provides a good fit to the intensity of commodity flows across commodity types, with or without directionality, and for both mass and value weighting schemes. Yet, Fig 3.5B shows that nodal mass flux decrease with spatial scale, as we would expect, even though a Gamma distribution provides a suitable model across scales.

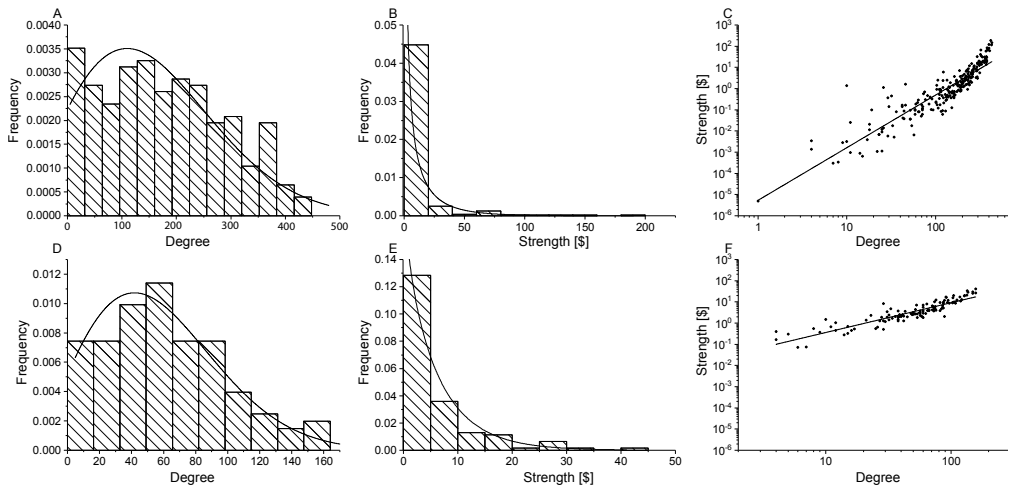

Figure 3.7: Network properties for global (top row) and national (bottom row) commodity flow networks by value [$]. Node degree distributions with generalized exponential binomial distributions fit to the data are shown in the first column (Panel A, D), node strength distributions with gamma distributions fit to the data are shown in the second column (Panels B, E), and power law relationships for node strength versus degree are shown in the third column (Panels C, F).

Here, we have shown that the Gamma distribution appropriately captures nodal strength distributions across all flow networks considered. This implies that it is a flexible statistical model for representing commodity flow systems. There are known properties of the Gamma distribution, which means that future efforts to model commodity flows may be able to benefit from these attributes. For example, the Gamma distribution has known reliability, lifetime, and hazard functions (*Agarwal and Kalla*, 1996). These statistical properties can be taken into account to model commodity flows in future research.

Relationship between connectivity and flows

We examine the relationship between node degree and strength for undirected food flows in Fig 3.2C, F, and I. In Fig 3.2C, F, and I s is plotted against k for all spatial scales. The straight line relationship in log-log scale indicates that there is scale invariance between mass flows and network connectivity. Specifically, a power law relationship between nodal mass flux and connectivity is evident across all spatial scales. Thus, there is a power law relationship between s and k food flows and it is consistent across village, national, and global food networks.

A linear relationship is fit to $\log(s)$ and $\log(k)$ such that:

$$\log(s) = a + b\log(k) \tag{3.3}$$

The parameters of the power law relationship for undirected food and non-food flows are provided in Table 3.3. The statistical distribution parameters change with the scale of analysis, indicating that there is scale dependence, despite the fact that the power law exists across scales. The power law exponent is the highest for global trade (slope = 2.7; see Table 3.3), but is similar for national and village scales (slope = 1.5 and 1.6, respectively). The power law relationship is less clear for non-food flows. Note that the points exhibit more scatter in Fig 3.3 than in Fig 3.2. Likewise, the exponents are consistently smaller for non-food flows than for food flows. Again, this indicates that food and non-food flow networks exhibit different properties which may be due to their underlying unique attributes.

What are the implications of a power law relationship for node strength versus degree? Particularly for global trade, the high b value indicates that there is a strong relationship between the mass that each nation trades and its number of trade partners. In other words, the node strength grows faster than node degree, so the more trade connections a country has, the much more it is able to participate in the exchange of commodity mass. This relationship occurs in a highly nonlinear way. In this way, shocks to trade relationships may prove highly disruptive to national access to the mass of food commodities, unless trade patterns are allowed to adapt. This is another statistical attribute that future efforts to model food flows may endeavor to incorporate.

Network clustering and centrality

The clustering coefficient enables us to evaluate the tendency of nodes in the network to form tightly connected groups. In Fig 3.5D, it is clear that node clustering decreases with spatial scale. In other words, nations are more likely to form 'cliques' than are households in a village (*Costa et al.*, 2007). However, node clustering decreases in a much less consistent manner than it does for degree and strength. This can be seen by the fact that the whiskers in the box-whisker plot overlap for clustering. Scale dependency in network parameters indicated by Fig 3.5D likely arises as a result of the aggregation process in food fluxes from smaller to larger scales of analysis.

Fig 3.8 presents the relationship between betweenness centrality (B) and degree (k) for food fluxes by spatial scale. Core nodes are those with both high k and B values. A core has been shown for global (*Ercsey-Ravasz et al.*, 2012) and national (*Lin et al.*, In Review) food flows. We confirm that this relationship exists for both undirected and directed food and non-food commodities (refer to Fig 3.8), although a core group of nodes is more pronounced for food networks.

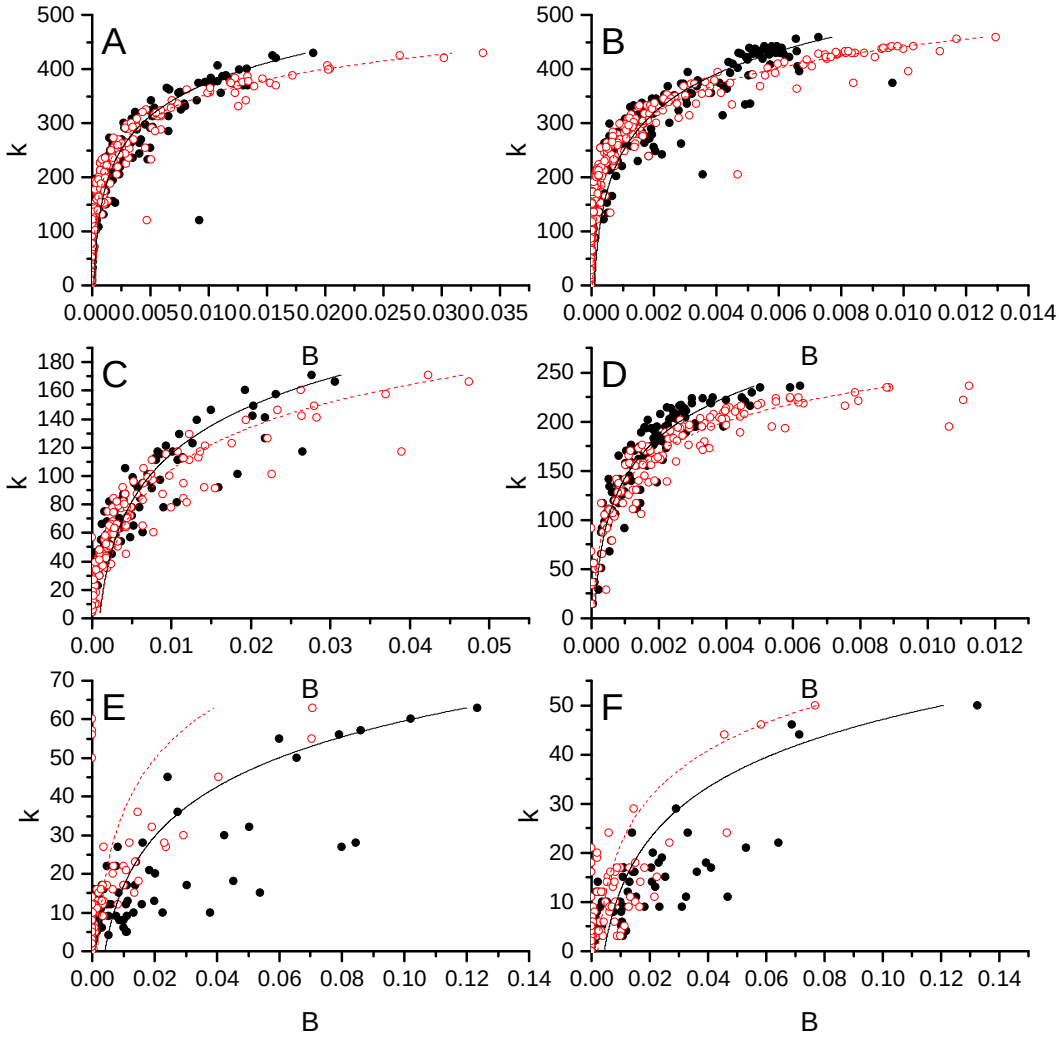

Figure 3.8: Node degree (k) versus betweenness centrality (B) for food and non-food commodity flow systems. The global scale is shown in the top row (Panels A, B), national scale is shown in the middle row (Panels C, B), and village scale is shown in the bottom row (Panels E, F). Food commodities are shown in the first column (Panels A, C, E) and non-food commodities are shown in the second column (Panels B, D, F). Red points show directed B and black points show undirected B. Note axes scales differ across panels.

We also show that a core group of nodes exists at the smallest spatial scale. In fact, core households at the village scale exhibit the highest centrality of all food flows networks. Note that the scale on the x-axis is an order of magnitude larger for the village scale than for the global scale. Direction is less important to node centrality at the village scale than it is at the national and global scales (note the black and red lines flip in Fig 3.8E, F). This indicates that some households are more instrumental to the structure and functioning of

village scale food exchanges than are countries at the global scale. The network core exists for all Alaskan villages, as shown in Fig 3.9.

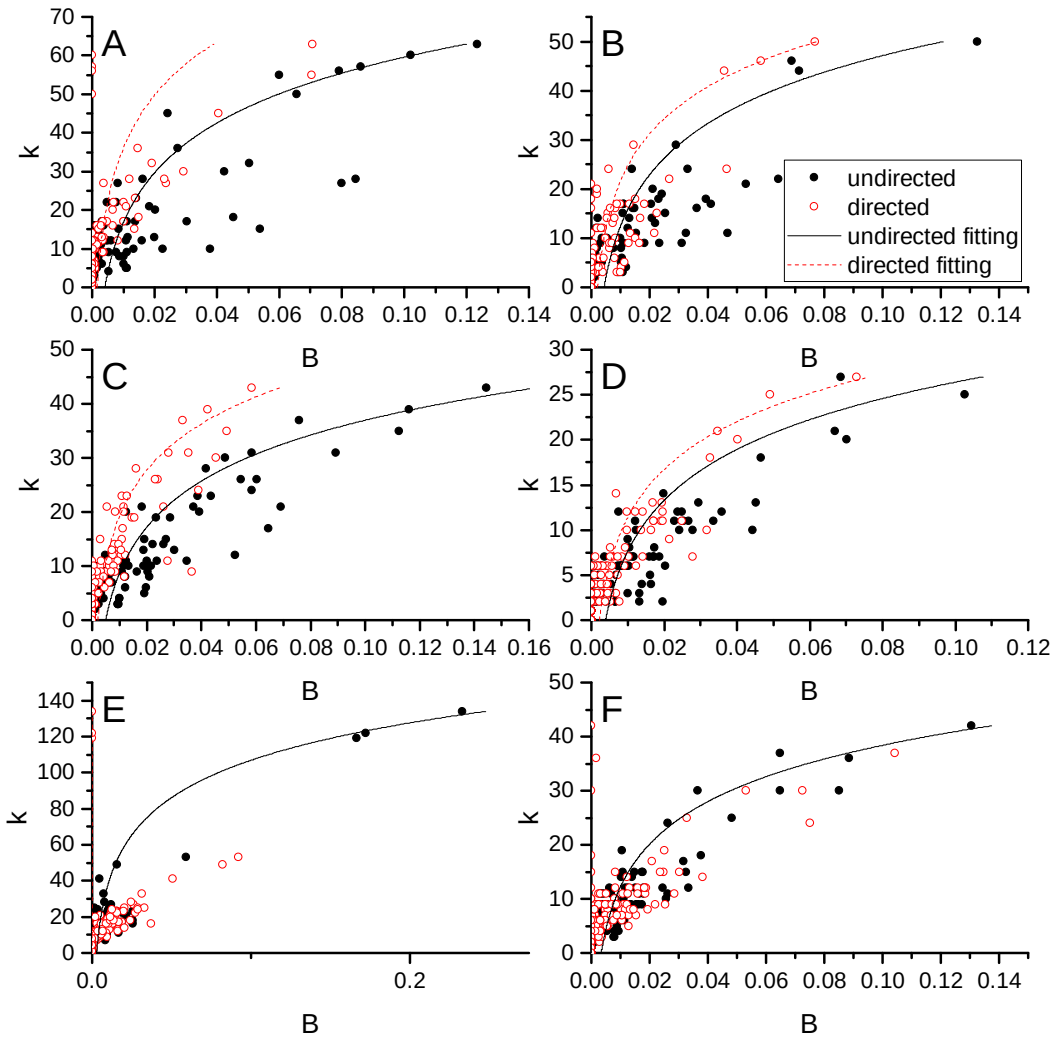

Figure 3.9: Node degree (k) versus betweenness centrality (B) for food and non-food village commodity flow systems. Kaktovic is shown in the top row (Panels A, B, C), Venetie is shown in the middle row (Panels D, E, F), and Wainwright is shown in the bottom row (Panels G, H, I). Food commodities are shown in the first column (Panels A, C, E) and non-food commodities are shown in the second column (Panels B, D, F).

Thus, the B versus k relationship is consistent across scales (shown in Fig 3.8). This indicates that a core group of nodes exists across spatial scales. Yet, scale dependence also exists since the statistical attributes of B vary with spatial scale. This result is shown in Fig 3.5C, in which B increases with decreasing spatial scale. At the village scale, B is more

concentrated amongst some households that are dramatic outliers. These B outliers are not evident in the global and national scales. So, some households in the village scale are more central to its food flow network than any FAF areas are to the national network or countries are to the global network. In this way, the village is more vulnerable to the removal of its core households.

3.4 Conclusion

Food flow networks may prove to be an important adaptation measure to cope with future climate and economic shocks. For example, if extreme climate events increase in frequency as projected under a changing climate, the ability to transfer commodities in both space and time may help those production locations that experience shocks to maintain consumption by importing from external sources. As such, it is essential to understand the scaling properties of food commodity flow networks so that we can understand how to model food flows and evaluate opportunities and roadblocks to the spatial and temporal redistribution of goods. Information on the scaling properties of food flow networks will also enable prediction of flows for the many locations and settings for which food transfer data is not available.

We have examined the empirical network structure of food commodity exchanges across the full spectrum of spatial scales. Village scale donations of food between household is the smallest spatial scale at which social food fluxes can occur; global scale international food trade between countries represents the largest possible spatial domain. Empirical evidence suggests that both scale dependent and invariant properties exist for food flow networks. Network parameters such as mean node connectivity, mass flux, and centrality vary with spatial scale, likely due to the aggregation process of food fluxes. Yet, we find that the statistical distribution functions of node connectivity and mass transfers are invariant across scales. Likewise, the relationship between node connectivity and mass flux exhibits a power law relationship for each spatial domain. These relationships hold for commodity fluxes weighted in both mass [kg] and value [$] units and for both undirected and directed networks. However, non-food commodities are not well fit by the same statistical distributions across spatial scales. This highlights that there are unique attributes of food transfers that lead them to have network properties that are distinct to non-food.

The network structures of food flow systems provide a signature of their vulnerability and resiliency to disturbance. Extensive research has explored the implications of certain network structures for vulnerability and resiliency. For example, networks with a power law node degree distribution have been shown to be vulnerable to targeted attack, but resilient to random attack. Future research is needed to explore the implications of the statistical network distributions of food flows presented here. Scale invariant properties indicate that

similar governing mechanisms are likely influencing food flows across scales, despite the fact that these systems are typically thought to arise from starkly different generative processes. We hypothesize that universal signatures of human behavior may lead to the similarities in food network statistical distributions across scales. For example, the human tendency for risk-sharing and cooperation may be an important mechanism generating the emergent food exchange patterns. Future research can build upon this work by modeling network formation processes and estimating food flows at resolutions lacking data.

CHAPTER 4

FOOD FLOWS BETWEEN COUNTIES IN THE UNITED STATES

4.1 Introduction

Most food security research focuses on increasing production (*Lobell et al.*, 2011; *Long et al.*, 2015; *Liang et al.*, 2017), but distribution through complex supply chains is even more critical to food security (*Ercsey-Ravasz et al.*, 2012; *Konar et al.*, 2018). In fact, the world already has 1.5 times of enough food to feed the whole population (*Holt-Giménez et al.*, 2012). This abundance can feed 10 billion people, which is the projected peak population in 2050. However, 1/7 of the world population is currently hungry (*Foundation*, 2018), while 1/3 of food is wasted. Food security is a problem even in the relatively affluent United States, which is the largest producers of agricultural crops in the world. Yet, according to *Feed America* (2018a), 1 in 8 Americans struggles with hunger. Even in the major food producinf states, like Illinois and Iowa, there is about 14% children in food insecurity (*Feed America*, 2018b). Hunger is not a scarcity problem, but a distribution problem, which means food flows are important to consider and understand. Much research attention has been devoted to improving food production, yet a comparable effort has not been given to food flow networks. Even in the USA, the most data-abundant country, the finest resolution of food flow is at a relatively coarse spatial scale (i.e. the FAF spatial units, see below). Given such a coarse resolution, it is difficult to understand the presence and absence of food flows in the USA, not to mention developing countries where data is even more deficient. Understanding food security really needs high resolution information on food flow network. Similarly, potential risks to the food flow network, such as potential water hazards and security threats, are difficult to quantify due to the limited spatial resolution of food flows.

Spatially detailed food flow estimates would improve our understanding of food supply chain vulnerabilities and enable spatially detailed footprint assessments. Global food trade has been shown to be vulnerable to disruptions, such as drought and food contamination (*Puma et al.*, 2015; *Ercsey-Ravasz et al.*, 2012). High-resolution understanding of food flows would enable a more accurate assessment of their vulnerabilities to disruption. Vulnerabilities in the food flow network within the United States have been evaluated at

a relatively coarse spatial scale (*Lin et al.*, In Review). A more spatially detailed understanding of food flows would help us to pinpoint critical locations and infrastructure in the national food supply chain (*Xu et al.*, 2012). Additionally, spatially resolved food flow estimates would advance life-cycle and footprint assessments of agricultural production and food consumption. Agricultural production is the dominant way in which people impact the environment (*Vitousek et al.*, 1997; *Tilman*, 1999; *Foley et al.*, 2005). Agriculture alters biogeochemical cycling (*Bouwman et al.*, 2013), impacts land cover (*Hansen et al.*, 2010), requires vast quantities of water resources (*Hoekstra and Mekonnen*, 2012; *Shen et al.*, 2018), and is responsible for significant carbon emissions (*Melissa et al.*, 2017), among other impacts. It is critical to spatially resolve food supply chains in order to assess their environmental footprint since the impact of consumer food goods are strongly coupled to the location of production (*Weber and Matthews*, 2008).

The United States is a key country in the global food system (*Xu et al.*, 2011). The U.S. produces over 30% of the world's corn and over 50% of the world's soybeans (*USDA*, 2013). The U.S. also accounts for large shares of the world export market for several staples: about 60% for corn, 40% for soybeans, 25% for wheat, and 70% for sorghum (*USDA*, 2013), making the U.S. an important contributor to global grain supplies (*FAO*, 2013). The ability to grow and transport agricultural products enables the U.S. to provide both domestic and global food security (*Lin et al.*, In Review). The U.S. is able to maintain its role as a key agricultural producer, consumer, and trade power largely due to its supporting water resources and food distribution infrastructure (*Marston et al.*, 2018; *Rushforth and Ruddell*, 2018). Supply chains in the U.S. are also responsible for a large national carbon (*Weber and Matthews*, 2008; *Cuéllar and Webber*, 2010; *Liang et al.*, 2016), water (*Dang et al.*, 2015; *Melissa et al.*, 2017; *Wang et al.*, 2017), and chemical pollution footprint (*Nesheim et al.*, 2015).

Data on subnational food flows is available within the United States at a coarse spatial resolution. This availability of subnational food flow information is a major reason for our selection of the U.S. for this work. The U.S. Census Bureau and the Bureau of Transportation Statistics produce the Commodity Flow Survey (CFS) every 5 years (ending in '2' and '7'). The CFS provides data on the movement of commodities in the United States, including their value, weight, and mode of transportation, as well as the origin and destination of shipments from manufacturing, mining, wholesale, and selected retail and services establishments. The Freight Analysis Framework (FAF) builds on the CFS data to provide data on freight movement between the 132 FAF zones of the U.S. (see Fig 4.1A) (*Oak Ridge National Laboratory*, 2015).

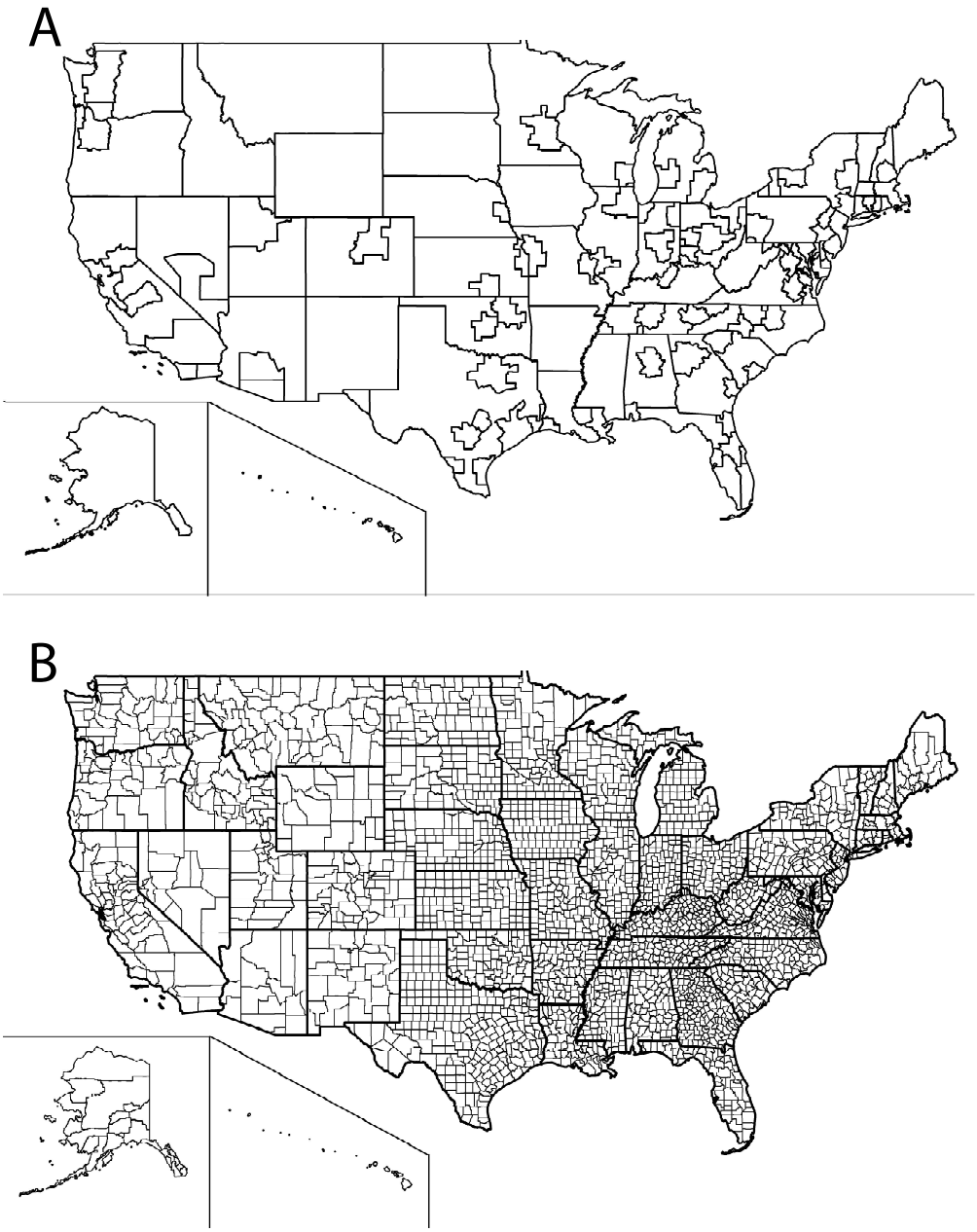

Figure 4.1: Maps of political boundaries within the United States. (A) Map of FAF zones. (B) Map of the counties of the United States.

FAF reports flows of coarse food commodity classes (see Table 4.1). Census information on food flows within the U.S. has been used to evaluate the network properties of the national food supply chain at the FAF zone spatial resolution (*Lin et al.*, In Review; *Konar et al.*, 2018). More spatially detailed food flows would enable a better understanding of the vulnerabilities and resiliencies within the U.S. food supply chain.

Table 4.1: List of Standard Classification of Transported Goods (SCTG) food categories included in this study.

SCTG	Model
1	Animals and Fish (live)
2	Cereal Grains (includes seed)
3	Agricultural Products (excludes Animal Feed, Cereal Grains, and Forage Products)
4	Animal Feed, Eggs, Honey, and Other Products of Animal Origin
5	Meat, Poultry, Fish, Seafood, and Their Preparations
6	Milled Grain Products and Preparations, and Bakery Products
7	Other Prepared Foodstuffs, Fats and Oils

Our work contributes to the recent literature that models food flows. A few recent papers have modeled detailed food flows (*Smith et al.*, 2017; *Venkatramanan et al.*, 2017). (*Venkatramanan et al.*, 2017) present a data-driven approach to estimate food flows between markets in Nepal in order to evaluate their propensity to spread pests. (*Smith et al.*, 2017) use a transportation optimization model to estimate corn flows between U.S. counties. Our approach is related but distinct. As in the existing literature, we use food production and consumption statistics in conjunction with a linear programming framework that minimizes transport cost. Now, we add some novel elements to the food modeling literature. We additionally constrain our food flows to have the same network properties as those of observed food flow networks (*Konar et al.*, 2018). Recent research has shown that global, subnational and village scale food flow networks share structural properties (*Konar et al.*, 2018). Specifically, all food flow networks exhibit a generalized exponential node connectivity distribution, Gamma mass flux distribution, and power law relationship between node connectivity and mass flux (*Konar et al.*, 2018). We utilize this empirical understanding in our novel methodological framework. Additionally, we incorporate the principle of mass balance to our model. Specifically, we require that the food flows from counties within a FAF zone sum to the food flows from that FAF zone. Both of these novel aspects enhance the realism of our approach.

The goal of this chapter is to develop a methodological framework to estimate food flows in locations without empirical flow data. We use this novel methodology to estimate food flows between all counties and county equivalents in the United States. There are 3,142 counties and county equivalents in the United States: 3,007 counties, 64 parishes, 19 organized boroughs, 10 census areas, 41 independent cities, and the District of Columbia. For the rest of this paper, we refer to these 'counties and county equivalents' simply as 'counties' for short. The major question that we address is: What are the food flows between counties within the United States? To answer this question, we develop the Food Flow Model. We detail our input data and the Food Flow Model algorithm in Section 3.2. We discuss our

results in Section 3.3. We conclude in Section 3.4.

4.2 Methods

4.2.1 Input data

We obtain two major types of data for this study. First, we obtain data on agricultural and food commodity transfers between FAF zones in the United States. Second, we obtain statistical information on economic production within each U.S. county. The FAF4 database that we obtain is for the year 2012. We obtain all other data for the year 2012 to match FAF4. All data sources are detailed in Table 4.2.

Table 4.2: List of data sources used in this study

Name	Ref.	Data Description	Purpose
Commodity Flow Survey (CFS) Public Use Microdata	(*US Census Bureau*, 2015a)	Survey of business shipments within the United States. FAF is largely based off this dataset, though the scope of the CFS Microdata is not as broad as that of the FAF dataset. However, the CFS Microdata contains greater shipment detail, including the NAICS industry responsible for the shipment.	This dataset allowed for pairing of commodity transfers to specific industries.
Freight Analysis Framework (FAF) Version 4	(*Oak Ridge National Laboratory*, 2015)	Data detailing freight movement between 132 major metropolitan areas and remainder of states (i.e., FAF Zones), as well as 8 international import/export regions.	FAF commodity transfers are used to constrain county transfers. The sum of county transfers must equal that of the FAF Zone that they belong.

Table 4.2 (cont)

Name	Ref.	Data Description	Purpose
US Census Bureau 2012 Economic Census	(*US Census Bureau*, 2015b)	Provides county level economic data by industry, including employment and the value of industry output.	The Economic Census was used to determine production of processed agricultural goods and the total production output of all industries using agricultural goods as production inputs. These data were used in our gamma mixture hurdle model for link prediction and assigning flow strength.
US Department of Agriculture 2012 Census of Agriculture	(*US Department of Agriculture*, 2014)	Agricultural production data for each crop or livestock type at the county scale.	The Census of Agriculture was used to determine county level production values for each crop and livestock. These data were used in our gamma mixture hurdle model for link prediction and assigning flow strength.
Input-Output (I-O) Accounts Data	(*US Bureau of Economic Analysis*, 2014)	These data detail supply chain input requirements for each industry per unit of their output.	Direct requirement coefficients from the I-O accounts were multiplied by production data to determine the commodity input requirements of each industry, as well as end consumers. A county's total input requirement of a commodity across all industries and end consumers represents its total consumption of that good. This is used in our gamma mixture hurdle model for link prediction and assigning flow strength.

Table 4.2 (cont)

Name	Ref.	Data Description	Purpose
County-to-County Distance Matrix and Network Impedance	(*Oak Ridge National Laboratory*, 2011)	Matrix of distances and impedances between county centroids via different transportation methods.	The linear programming algorithm used this matrix to minimize transportation cost.
Personal Income	(*US Bureau of Economic Analysis*, 2017)	Personal income data per county.	When paired with the input-output data tables, this was used to help determine final consumer demand of different commodities within a county.
Port trade	(*US Census Bureau*, 2018)	Value [$] and mass [kg] trade data for international ports of the United States.	Trade data to/from these ports was used to better capture transit hubs in the Gamma mixture model.

Empirical agricultural and food commodity transfers come from the Freight Analysis Framework Version 4 (*Oak Ridge National Laboratory*, 2015). The FAF4 dataset utilizes data from numerous sources to provide an exhaustive description of subnational freight movement in the United States, as well as trade with major international regions. The Commodity Flow Survey (CFS) is foundational to the FAF4 dataset. Every 5 years (years ending in '2' and '7'), the CFS samples more than 100,000 establishments that ship freight domestically. Survey responses are aggregated to the corresponding FAF zone, commodity class, and across the 4 quarterly surveys administered during the year of record to protect the confidentiality of survey respondents. Freight shipments within the CFS and FAF dataset are grouped into 42 classes using the two-digit Standard Classification of Transported Goods (SCTG). Here, we are primarily interested in agriculture and food goods, which are represented by SCTG 01-07 (Table 4.1). We use the FAF4 commodity transfer database to develop regression models of commodity transfers between FAF zones that are then applied to the county spatial scale (see the following section for more details concerning the construction of the regression models). This approach assumes that the regression model is consistent across spatial scales. The FAF4 data is also used to constrain transfers within our county-to-county Food Flow Model. We utilize the principle of mass balance to ensure that counties located within a FAF zone do not exceed the mass flux reported at the FAF spatial scale.

The likelihood and mass flux of food transfers originating from a county is related to its production of these goods. County level production [$] for unprocessed agricultural commodities (SCTG 1-4) come from (*US Department of Agriculture*, 2014). Production values of each crop or livestock category originating within a county were aggregated to their corresponding SCTG code. Similarly, the county level production of processed agricultural and food goods (SCTG 5-7) were aggregated to their respective SCTG code as described below. Production data of processed agricultural goods originating from a specific NAICS food processing industry comes from (*US Census Bureau*, 2015b). Distance between all county pairs was obtained from (*Oak Ridge National Laboratory*, 2011) and represents the Euclidean distance between county centroids.

Production data is essential to determine the origin of food flows. Other statistical information is required to determine the destination of food flows. The 2012 CFS Public Use Microdata (*US Census Bureau*, 2015a) and the United States Bureau of Economic Analysis input-output accounts data (*US Bureau of Economic Analysis*, 2014) were used to statistically determine the production and attraction of food within our machine learning algorithm (see below). The CFS Microdata utilizes the same survey data as the CFS dataset but provides greater shipment detail than the standard CFS data. Importantly, one additional detail included in the CFS Microdata is the North American Industry Classification System (NAICS) code of the industry producing and shipping the good. This additional information enables us to relate the SCTG code of a transported commodity to the NAICS industry producing the commodity. Since the CFS Microdata does not provide a NAICS code for raw agricultural and food goods (SCTG 01-04), we manually matched the production of individual crops or livestock reported by (*US Department of Agriculture*, 2014) to the SCTG code to which it belongs (a listing of goods within each SCTG can be found at (*SCTG*, 2017)). The SCTG-NAICS crosswalk table we created (provided in the Supporting Information) was paired with input-output accounts data to determine an industry's use of each SCTG as input in its production process. Input-output tables show to what degree the production (output) of one industry is used as input to another industry. Using the crosswalk table we created, we aggregate industry output within the table to its corresponding SCTG code to match the FAF4 data set. This procedure allows us to restrict data used within our machine learning algorithm to variables that have been established as relevant to the production or consumption of each SCTG good. This ensures that our model maintains realism.

Some agricultural (*US Department of Agriculture*, 2014) and business production data (*US Census Bureau*, 2015b) are suppressed by the data collecting agency if their release may reveal information on an individual producer. Suppressed data records are not removed from the data set, but instead flagged, indicating there are limited producers within that

geographical area. Data suppression is more prevalent at the county spatial scale and among specialty producers. For example, artichoke production in Linn County, Oregon is flagged since reporting this data would reveal information specific to the only artichoke farmer in the county. When suppressed values arise in the data sets, the geographical and industry/product hierarchical structure of the data is exploited to estimate these suppressed values. The artichoke production of the sole farmer in Linn County, for example, was estimated by subtracting the sum of all artichoke production in Oregon counties from the state-level production value provided by (*US Department of Agriculture*, 2014). The difference between the state total and the sum of all counties is uniformly distributed amongst all Oregon counties with suppressed artichoke production records. Industrial production records have other data fields that can help us further refine our estimates of suppressed production values. (*US Census Bureau*, 2015b) provides employment records for each industry within a county, which can be used to help estimate suppressed production output. Relationships between production output and employment were established for every industry based on the large number of records where both values were provided. This allowed us to estimate production for counties with limited industrial activity. While similar approaches have been applied in the literature (*Isserman and Westervelt*, 2006; *Smith et al.*, 2017; *Marston et al.*, 2018), our study would nonetheless benefit from a complete data set.

Port trade data is retrieved from the Census Bureau USA Trade database (*US Census Bureau*, 2018) for the year 2012. The values [$] and mass [kg] for both sea and air ports are provided (*US Census Bureau*, 2018). Value flows were ultimately used due to significantly more data availability as compared to mass. While land ports are not specifically mentioned, many of the reported ports are US Customs and Border Patrol crossings on US land borders (such as along the Northern borders of North Dakota and Montana), implying that land ports are included in the database. Commodities in the port trade database are reported using the HS coding system. For consistency with FAF flow data, a crosswalk was created to convert from HS to SCTG codes. The Python geocoder library (*python*, 2018) was then used to determine latitude and longitude coordinates for each port. Some ports, such as Low Value (Port), did not have locations and were consequently removed. A spatial join was finally used to determine which county each port is in, resulting in 331 ports in 228 counties contributing inflows and outflows of SCTGs 1 through 7 in the US.

Refer to Table 4.3 for the list of variables used.

Table 4.3: List of model variables. Note that the variables in the top portion of the table represent matrices (entry for each link), while the variables in the bottom portion of the table are vectors (entry for each county).

Variable	Description
F	Food flows in mass [kg] for all county pairs
A	Adjacency matrix of connectivity for all county pairs
D	Distance between all counties
GDP	Gross Domestic Product [$]
POP	Population
P	Production [tons]
IND	Sum of industrial products utilizing a particular SCTG as input
LIVE	Production of all livestock
A1	Animal slaughtering and processing
A2	All other food manufacturing
A3	All other miscellaneous chemical product and preparation manufacturing
B1	Bakeries and tortilla manufacturing
C1	Cattle
C2	Chickens
C3	Coffee and tea manufacturing
D1	Dairy product (except dried or canned) merchant wholesalers
D2	Dry, condensed, and evaporated dairy product manufacturing
F1	Fruit and vegetable canning, pickling, and drying
F2	Food and beverage stores
F3	Fresh fruit and vegetable merchant wholesalers
F4	Frozen food manufacturing
F5	Fats and oils refining and blending
G1	Goats
G2	General merchandise stores
G3	Gasoline stations
G4	Grain and field bean merchant wholesalers
G5	Grain and oilseed milling
H1	Hogs
I1	Ice cream and frozen dessert manufacturing
M1	Meat and meat product merchant wholesalers
M2	Meat processed from carcasses
O1	Other animal food manufacturing
O2	Other food manufacturing
P1	Poultry processing
R1	Rendering and meat by product processing
R2	Poultry and poultry product merchant wholesalers
S1	Seafood product preparation and packaging
S2	Sugar and confectionery product manufacturing
S3	Snack food manufacturing
S4	Soft drink and ice manufacturing
S5	Sheep
S6	Seasoning and dressing manufacturing
S7	Support activities for transportation
T1	Turkeys
W1	Wholesale trade

4.2.2 Link existence estimation mathematical foundation

Introduction to generalized exponential distribution

Generalized exponential distribution is generated from three independent random Poisson processes. To understand how this distribution is generated, imagine there is location i and its counterpart. Location i has one independent component. Its counterpart also has one independent component. Besides, they share a third component. A breakdown of any of these three component will lead to the system to be nonoperational, thus stopping forming new links. The individual components in location i and its counterpart following Poisson shock processes, with fixed intensity rate. The shared component follows cumulative nonfatal Poisson shock process. This may reveal underlying mechanism for network degree. More explanation about the connection between distribution and fodd flow is in A.4.

Link existence estimation method

Model with non-Gaussian distributed observation is commonly known as generalized linear model (GLM) (*Faraway*, 2016). Ideally, link existence estimation could be adapted generalized exponential distribution within generalized linear model. However, it is hard to deduct from $P(link_{ij} = generalized \quad exponential)$ to $P(link_{ij}|X_{ij})$. Since it is not a popular distribution, there is no existing applicable solution for such a GLM problem. Logistic regression is a widely accepted method for link existence estimation problems. Here it is taken as an simplified alternative to estimate link existence between any pair of locations by assuming $P(link_{ij}|X_{ij}) = f(X_{ij})$.

Random process assumption with logistic regression As logistic regression is adopted to estimate link existence, effectively the linkage between any pair of locations is modeled by a Bernoulli trial, $P(link_{ij}|X_{ij}) = f(X_{ij})$, where we assume between any pair of locations, a Bernoulli trial is conducted, whose success probability is decided by environmental variables, such as distance, population, commodity production and etc. Notice that this would not generate Binomial degree distribution in the network because the success probabilities varies across different pairs and consequent network degree distribution is decided by the environmental variables. No guarantee that this assumption will lead to $P(link_{ij})$ following generalized exponential distribution. If the resulting network degree distribution is not generalized exponential distribution, this assumption becomes problematic.

4.2.3 Flow strength estimation mathematical foundation

Introduction to Gamma distribution

To better explain Gamma distribution, Bernoulli trial process, Poisson process and negative binomial distribution are introduced. The purpose here is not to provide exact definition, but to help intuitively understand these concepts.

Introduction to Bernoulli trial process and Poisson process In probability theory, an experiment/trial is defined as a procedure that can be repeated indefinitely and has a definite set of outcomes (sample space). A Bernoulli trial is a special experiment where there are only two potential outcomes, each with a fixed chance of happening. A sequence of such independent Bernoulli trials is defined as Bernoulli trial process.
Poisson process is the continuous analogue of Bernoulli trial process. Instead of discrete number of trials, Poisson process has infinite trials.
For an example, assuming a phone has equal probability of receiving a call at any time in 5 mins and each phone call is independent, then the results from 5 experiments in 0-1 min, 1-2 min, 2-3 min, 3-4 min and 4-5 min compose a Bernoulli trial process, while the infinite trials happens at any time within these 5 mins constitute a Poisson process.

Introduction to negative binomial distribution and Gamma distribution In probability theory, the negative binomial distribution is a discrete probability distribution for the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures/successes occur. And Gamma distribution is the continuous analogue of it.
Let's assume a visitor is throwing bananas at a monkey (this visitor has infinite amount of bananas and the monkey catching bananas has constant probability and each catch is independent). All the results from trials with the monkey catching the banana constitute a Bernoulli process and the total number of bananas the visitor need to throw before this monkey gets exactly 5 bananas will follow a negative binomial distribution. If the visitor is throwing any unit weight of banana at the monkey , the monkey has a constant probability density to catch any unit of banana and each catch is independent, all the trials will constitute a Poisson process, and the total weight of banana the visitor need to throw at the monkey before this monkey catches 1 kg banana follows Gamma distribution.

Poisson process assumption on food flow

We have listed some consistent network properties in chapter 3. Further, we found $flow_{ij}$ also follow Gamma distribution with high $adj - R^2$ ($> 97\%$). In attempt to unveil the underlying mechanism causing these properties, after trials and fails, we found Poisson process in food flow explains Gamma distribution in $flow_{ij}$. In the following discussion, we will illus-

trate a potential mechanism that leads to this $flow_{ij} \sim Gamma$ property. And in the result section, through modeling and simulation, we will demonstrate this mechanism will not only lead to this network flow strength distribution property, but also the other properties I have concluded from chapter 3.

We assume the process of transferring a unit of weight of food from location i to j as an *experiment*, which has exactly two potential outcomes, *success* (this transfer of commodity fulfills both the need of location i to export as well as need in location j to import) and *failure* (this transfer doesn't fulfill either location i's need to export or location j's need to import) and the success probability density is a constant, $\frac{1}{\theta}$. Then the total commodity weight transferring from i to j is a Poisson process with success probability density $\frac{1}{\theta}$. Based on the relationship between Poisson process and Gamma distribution, from our assumption we can deduct that

$$P(flow_{ij}|X_{ij}) = Gamma(k(X_{ij}), \theta)$$

where X_{ij} is environmental variables in i and j, $k(X_{ij})$ is expected amount of weight need to be successfully transfered (it is a function of environmental variables), and $\frac{1}{\theta}$ is the success rate (probability of success per unit weight of commodity).
This equation means for any pair of locations, i and j, as their environmental variables (X_{ij}) are given, the commodity flow strength between them follows a Gamma distribution, where *shape* ($k(X_{ij})$) is a function of the environmental variables, and *scale* (θ) is a constant.
To help interpret this equation, imagine there are two villages, i and j. i is closer to a banana production area or i is a banana producer, while j have some children love bananas. There are many environmental variables, X_{ij}, including distance (it impacts cost to transport bananas from village i to village j), number of children like bananas, abundance of bananas in village i, and local economy in village j (it impacts their purchasing power), together decide the quantity of required bananas to be consumed by these children, *shape* ($k(X_{ij})$, expected consumption quantity). And there is a success rate, $\frac{1}{\theta}$, with which for each unit weight of banana transported, a coin with a success rate of $\frac{1}{\theta}$ will be tossed. When the coin heads, the unit weight of banana will be consumed by children. Otherwise, the banana will not be consumed by children in village j,which might be staled during transportation, or re-transported to a third village, or etc. This rate is assumed to be a constant, indicating that for any other pair of villages k and h, their coin bias (success rate) will be the same. In the next session, we will prove if there are many such villages, the overall flow strength distribution across the whole network will also follow gamma distribution.

Deduction of network commodity flow distribution

$$
\begin{aligned}
P(flow_{ij}) = \\
Law \quad of \quad total \quad probability \Rightarrow &= \sum_{ij} P(flow_{ij}|X_{ij}) \times P(X_{ij}) \\
P(X_{ij}) = 1/N, N = number \quad of \quad links \Rightarrow &= \sum_{ij} P(flow_{ij}|X_{ij}) \times \frac{1}{N} \\
Assume \quad P(flow_{ij}|X_{ij}) = Gamma(k(X_{ij}), \theta) \Rightarrow &= \sum_{ij} Gamma(k(X_{ij}), \theta) \times \frac{1}{N} \\
Scaling \quad Property \quad of \quad Gamma \quad Distribution \Rightarrow &= \sum_{ij} Gamma(k(X_{ij}) \times \frac{1}{N}, \theta) \\
Summation \quad Property \quad of \quad Gamma \quad Distribution \Rightarrow &= Gamma(\sum_{ij} k(X_{ij}) \times \frac{1}{N}, \theta) \\
Let \quad K = \sum_{ij} k(X_{ij}) \Rightarrow &Gamma(K, \theta)
\end{aligned}
\tag{4.1}
$$

Thus, given the assumption that commodity transportation between any two locations is a Poisson process with a constant success rate, the commodity flows across the whole network would follow a Gamma distribution with the same success rate.

Difference between conditional flow strength distribution and network flow strength distribution

Sometimes, the relationship between $P(flow_{ij})$ and $P(flow_{ij}|X_{ij})$ can be confusing. One question might arise: "Gamma distribution is assumed and Gamma distribution is produced in the result. Why is this interesting? Are we proving what we assume?" There are three reasons. First, not any assumption of the distribution $P(flow_{ij}|X_{ij})$ leads to the same distribution $P(flow_{ij})$. Imagine a network, where $P(flow_{ij}|X_{ij})$ *lognormal* distribution. In this network, we won't have *Scaling property* or *Summation property*, so the resulting $P(flow_{ij})$ will depends on the distribution of X_{ij} and can be very different from *lognormal* distribution. Second, Gamma distribution has well-defined semantics that interprets the commodity flow process well. As discussed above, random variable, total quantity of item needs to be experimented until achieving a certain successes follows Gamma distribution. In this paper, it is the total weight of commodity needs to be delivered from one location to another until required successful delivery happens. Third, based on this assumption, the model has good performance fitting the real data at FAF scale. Furthermore the output from this model at county scale reproduces the consistent patterns more than Gamma strength

distribution we have observed in the previous chapter.

Flow strength estimation method

Based on the discussion above, it is concluded that if we can prove $P(flow_{ij}|X_{ij}) = Gamma(k(X_{ij}, \theta))$, we would be able to explain the underlying mechanism forming the Gamma distribution in $P(f_{ij})$. This section will demonstrate this equation's establishment by showing the goodness-of-fit of this model to FAF level commodity flow data.

Relationship between modeling and distribution Modeling is a process to estimate the unknown parameters in a model so that the model can best describe the data. Maximum likelihood estimator(MLE) is a common method to estimate these parameters.
With the conditional distribution assumption, we are able to provide MLE of our problem:

$$\begin{aligned} \mathcal{L}(\beta, \theta; flow_{ij}, X_{ij}) &= \prod_{i!=j} P(flow_{ij}|X_{ij}) \\ &= \prod_{i!=j} Gamma(h(\beta, X_{i,j}), \theta)) \end{aligned} \tag{4.2}$$

$$\begin{aligned} \hat{\beta}, \hat{\theta} =& argmax_{\beta,\theta}(\prod_{i!=j} Gamma(h(\beta, X_{i,j}), \theta))) \\ =& \prod_{i!=j} \frac{1}{\Gamma(h(\beta, X_{ij}))\theta^{h(\beta,X_{ij})}} flow_{ij}^{h(\beta,X_{ij})-1} e^{-\frac{flow_{ij}}{\theta}} \end{aligned} \tag{4.3}$$

Here $h(\beta, X_{ij})$ is a function of variable X_{ij} and its coefficients β, Γ is gamma function (*Artin*, 2015). The modeling process is to estimate β and θ given all flow data and environmental variables in these locations. Iteratively reweighted least squares algorithm (*Holland and Welsch*, 1977) is used as the numerical method to solve for these parameters.

4.2.4 Gamma hurdle model

To summarize, logistic regression would be utilized to estimate existence of a link/flow between any two locations and Gamma regression would fit flow strength for an existing flow. Combination of these two methods is commonly named Gamma hurdle model (*Burton et al.*, 2017).

Modeling link existence

Variables selection and regression Variable selection is carried out with balance-weighted (*Bohannon*, 1995) logistic regression with L1 penalty (*Tibshirani*, 1996). After variable selection with L1 penalty, T test (*Ruxton*, 2006) is conducted to check significance of each variable. Variables not showing significance are removed. This process is conducted for both original variables and log transformed variables. Because the 10-fold cross validation Area Under

Curve (AUC) (*Hanley and McNeil*, 1982) in log transformed model outperforms original for all SCTGs, the final regression model is generated with log transformed selected variables. The resulting logistic regression model for each SCTG group is provided in Table 4.4.

Table 4.4: Logistic regression model for each Standard Classification of Transported Goods (SCTG) food category. These regression models are based upon the gravity model of trade and predict if a link exists between all county pairs. A is the adjacency matrix for all county pairs. The A matrix has a value of '1' if a link exists between a county pair and a value of '0' if there is no link. D is distance, GDP is Gross Domestic Product [$], POP is population, and P is production [tons]. Subscript o indicates the variable of the origin county, while d indicates the variable of the destination county. Refer to Table 4.3 for the full list of variable names.

SCTG	Model
1	p4cm—logit(A1) = - 2.60 log(D) - 0.68 log(GDP_o) + 0.43 log(GDP_d) + 1.44 log(POP_o) + 0.03 log($R1_d$) + 0.04 log($S1_d$) + 0.14 log($LIVE_o$) + 18.32
2	logit(A2) = - 2.13006975 log(D) + 0.58 log(GDP_o) + 0.34 log(GDP_d) - 0.59 log(POP_o) + 0.37 log(P_o) + 0.03 log($O1_d$) + 0.14 log($C1_d$) + 0.10 log($G1_d$) + 6.63
3	logit(A3) = - 0.65 log(D) + 0.51 log(POP_o) + 0.56 log(POP_d) + 0.42 log(P_o) + 0.03 log($F4_d$) + 0.04 log($S6_d$) + 0.00 log($F1_d$) - 0.30
4	logit(A4) = - 1.71 log(D) + 0.22 log(GDP_o) + 0.48 log(POP_d) + 0.43 log(P_o) + 0.05 log($I1_d$) + 0.03 log($O1_d$) + 0.09 log($G1_d$) + 0.08 log($S5_d$) + 3.55
5	logit(A5) = -1.19 log(D) + 0.21 log(GDP_d) + 0.43 log(POP_d) + 0.19 log($F2_d$) + 0.18 log($W1_d$) + 0.06 log($A1_o$) + 0.09 log($M1_o$) + 0.05 log($R2_o$) + 0.06 log($S1_o$) + 0.32 log(IND_o) + 0.13 log($C1_o$) - 0.22 log($G1_o$) + 0.18 log($H1_o$) - 9.56
6	logit(A6) = -1.13 log(D) - 0.04 log(GDP_o) + 0.06 log(GDP_d) + 0.42 log(POP_d) + 0.39 log($G2_d$) + 0.02 log($S3_d$) + 0.59 log($B1_o$) + 0.32 log(IND_o) + 0.08 log($G1_d$) - 9.72
7	logit(A7) = - 1.40 log(D) - 0.72 log(POP_o) + 0.01 log($A2_d$) + 0.02 log($C3_d$) + 0.26 log($F2_d$) + 0.25 log($G3_d$) + 0.31 log($G2_d$) + 0.03 log($O1_d$) + 0.02 log($S4_d$) + 0.14 log($W1_d$) + 0.02 log($A3_o$) + 0.06 log($D1_o$) + 0.07 log($F3_o$) + 0.02 log($O2_o$) + 0.02 log($S4_o$) + 0.03 log($S2_o$) + 0.67 log(IND_o) + 0.02 log($C2_o$) + 0.15 log($S5_o$) + 0.04 log($T1_o$) + 0.74

Modeling flow strength

Least absolute shrinkage and selection operator(LASSO) (*Tibshirani*, 1996) variable selection and Gamma regression with log link function (*Ramsey and Schafer*, 2012) are conducted on location pairs where flow exists. Results with both original and log transformend input

variables are compared, and log transformed input variables have better performance. The Gamma regression model for each SCTG group is provided in Table 4.5.

Table 4.5: Gamma regression model for each Standard Classification of Transported Goods (SCTG) food category. These regression models are based upon the Gamma probability distribution and predict the weight of each link that exists between all county pairs. F is the weighted and directed matrix for all county pairs. The F matrix has a flow value for all predicted links in the adjacency matrix (A) (refer to Table 4.4). D is distance, GDP is Gross Domestic Product [$], POP is population, and P is production [tons]. Subscript o indicates the variable of the origin county, while d indicates the variable of the destination county. Refer to Table 4.3 for the full list of variable names.

SCTG	Model
1	$\ln(E(F1))$ = -0.61 log(D) + 0.11 log(GDP_o) + 0.05 log($M2_d$) + 0.08 log($P1_d$) + 0.10 log($R1_d$) + 0.21 log($H1_o$) + 0.17 log($LIVE_o$)
2	$\ln(E(F2))$ = -0.62 log(D) + 0.59 log(P_o) - 0.13 log(P_d) + 0.04 log($F1_d$) + 0.31 log($C1_d$)
3	$\ln(E(F3))$ = -1.30 log(D) + 0.23 log(POP_d) + 0.28 log(P_o) - 0.08 log(P_d) + 0.04 log($F1_d$) + 12.06
4	$\ln(E(F4))$ = -0.89 log(D) + 0.54 log(GDP_o) - 0.65 log(POP_o) + 0.46 log(P_o) - 0.20 log(P_d) + 0.03 log($O1_d$) + 0.38 log($C1_d$)
5	$\ln(E(F5))$ = -0.71 log(D) + 0.03 log($O1_d$) + 0.33 log(P_o) + 0.12 log($C1_o$) + 0.08 log($C2_o$) + 0.12 log($H1_o$)
6	$\ln(E(F6))$ = -0.77 log(D) + 0.03 log($F5_d$) + 0.35 log($G2_d$) + 0.04 log($S2_d$) + 0.08 log($S7_d$) + 0.12 log($G4_o$) + 0.05 log($G5_o$) + 0.12 log(P_o)
7	$\ln(E(F7))$ = -1.35 log(D) + 0.02 log($D2_d$) + 0.03 log($F5_d$) + 0.44 log($G3_d$) + 0.03 log($S4_d$) + 0.57 log(P_o) + 0.28 log($C1_o$)

Relationship to gravity model of trade In gravity model of trade (*Burger et al.*, 2009), $flow_{ij} = G\frac{X_i^a X_j^b}{distance^c}$, where G is a constant, X_i is independent variable like GDP, production, etc, and a, b, c are the coefficients. Taking log at both sides of the equation, we have $logflow_{ij} = logG + alogX_i + blogX_j - clog(distance)$. During regression, a normally distributed error ϵ is assumed, $log(flow_{ij}) = logG + alogX_i + blogX_j - clog(distance) + \epsilon$. Since the error is assumed to be Gaussian, a ordinary linear regression is justified to be conducted to estimate the coefficients in this equation. In our gamma regression with log link function, the equation is $log(flow_{ij}) = constant + alogX_i + blogX_j + clog(distance) + \delta$, where δ follow Gamma distribution. As it shows, they are the same structure except that error in our model is assumed to follow Gamma distribution instead of normal distribution, which is evidenced by patterns we demonstrated. Sharing the same structure makes our result is more comparable to gravity model of trade and in the future work, a comparative study of this result againt existing literature on gravity model of trade is made possible.
For most food commodity groups about 5% of the flows exceed the upper bound of the 95% confidence interval of our Gamma regression model. These outliers are major transportation hubs within the US.

4.2.5 Global sensitivity and uncertainty analysis

Global sensitivity and uncertainty analyses (GSUA) (*Saltelli et al.*, 1999) is the study of the impact of uncertainty in input sources to variation in the output of a physical or mathematical system. Through GSUA, impact of each variable on the commodity flow existence of strength will be demonstrated. And such an understanding has strong potential management strategy implications. By understanding which variable has the greatest impact on the flow strength, to increase commodity connections between these locations, potential management strategies can be formed.

Specifically, we now perform GSUA with the Fourier Amplitude Sensitivity Test (FAST) (*Saltelli et al.*, 1999). FAST is a GSUA method. One most prevailing GSUA was conducted by Morris Method (*Morris*, 1991) followed by Sobol's method (*Sobol*, 2001). Morris method is a pre-screening method to reduce the number of variables for global sensitivity analysis. We don't have many variables in our final models(< 20), so pre-screening Morris method is not required. Both FAST and Sobol's methods are GSUA methods. They are used to solve the same problem, and both approaches estimate "first-order" and "total-order" indices (*Sobol*, 2001), albeit using different sampling and computational methods. The main advantage of the FAST method is that it is robust for relatively small sample sizes (*Cukier et al.*, 1973). Additionally, the FAST method is computationally efficient (*Saltelli et al.*, 1999). In Bayesian notation, the total contribution of each input variable to the output, sometimes referred as total-effect index in Sobol's method is expressed as

$$Var_x[E(Y|X)]/Var(Y)$$

where X denotes input variable, Y denotes the output variable, and $Var_x[E(Y|X)]$ is the variance of $E(Y|X)$ over possible X values. This ratio represents the total contribution of input X on the output variance, including impact of X on output through interaction with other inputs.

In the analysis process, minimum value and maximum value of the random variable are taken as the bounds. Input random variables are assumed to follow lognormal distribution. We keep increase simulation iterations by 1000 until total-effect index's change between previous and current session is within 5% for any variable factor.

4.2.6 Flow estimation at county level

Downscale flow transfers at FAF scale to county scale Specifically, we use data on food transfers at the Freight Analysis Framework (FAF) zone spatial scale (refer to

Figure 4.1A) to estimate food transfers between counties within the United States (refer to Figure 4.1B). From Figure 4.1 it is clear that our goal requires the estimation of flows at a much finer spatial resolution (i.e. between all 3,142 county pairs) than that for which information is available (i.e. between 132 FAF zones). Since the number of directed paths is determined by $(n)(n-1)$, this means that our goal requires that we move from a system with 17,292 potential links (n=132 at FAF zone scale) to estimating 9,869,022 potential links (n=3,142 at county scale). As such, flow estimation quickly increases in complexity and computational demands as the number of nodes increases. In this way, our problem is distinct to most other spatial downscaling problems, in which a coarse spatial value is assigned to entities within its domain (see Fig 4.2A). Instead, we want to downscale *flows*, which requires estimating values (including zeros) between all node pairs (i.e. links) in our system. Fig 4.2 presents a conceptual framing of this challenge.

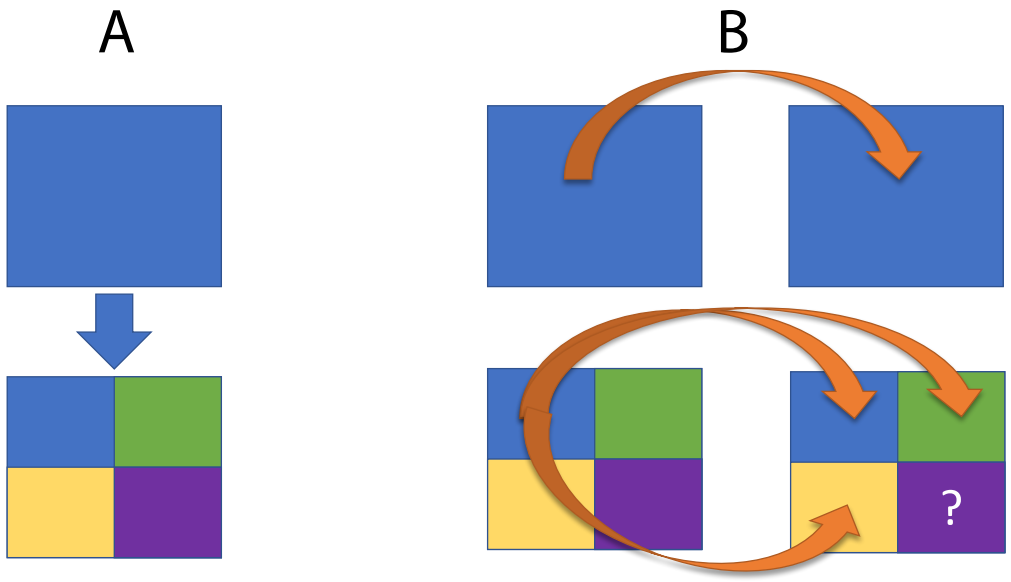

Figure 4.2: Conceptual schematic of the problem statement. The goal of the Food Flow Model is to downscale spatial flow data. This is more complex than traditional spatial downscaling (A), which typically aims to estimate attributes of sub-entities within a larger entity. Now, we aim to estimate connections and mass transfers between all node pairs (B). There are many more node pairs than spatial units which increases the entities to be estimated.

4.2.7 Flow estimation at county level method

Base on the Gamma hurdle model we developed in 4.2.4, we will estimate F in this section, which is a weighted and directed matrix of food flows between all county pairs (i.e. for all 9,869,022 potential links within the nation), where value at row i and column j is $flow_{ij}$. F provides flows from an origin (o) to a destination (d) county. Given this matrix, we will further validate both our assumptions and model by demonstrating the generated county level flow network showing consistent patterns as we have observed in the real commodity flows.

Figure 4.3 shows a process of estimating flow strength in county level network. In the first step, logistic regression is conducted to decide which links should exist among all county pairs. Once it decides which pairs have links between them, gamma regression is utilized to find the potential flow strength. Given estimated coefficients($\hat{\beta}$ and $\hat{\theta}$) and environmental variables in these two locations(x_{ij}), we have the flow strength distribution from i to j, $P(flow_{ij}|\beta = \hat{\beta}, \theta = \hat{\theta}, X_{ij} = x_{ij}) = Gamma(h(\hat{\beta}, x_{ij}), \hat{\theta}))$, which is the distribution we used to estimate flow strength in every link. Consider these potential flow strength as the capacity of pipelines, and distribute the food flow value (mass balance constraint) from FAF zone O to FAF zone D to these "pipes" with shortest travel distance fill first principle.

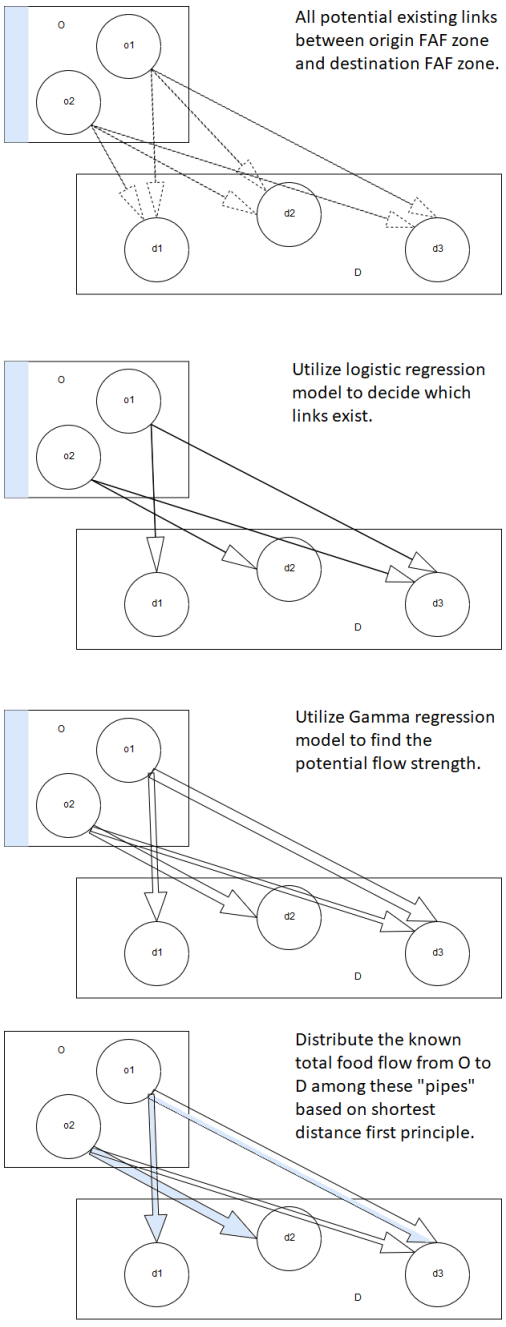

Figure 4.3: Flow estimation

4.3 Results and discussion

4.3.1 Logistic regression results

Logistic regression performance evaluation Table 4.6 lists logistic regression performance estimated with 10-fold cross validation (*Kohavi et al.*, 1995) AUC, R^2, predicted network density, and actual network density for SCTG 1 to 7 trained with FAF area food flow network. Cross validation is a method to avoid performance evaluation randomness by repeatedly and randomly cut training data and testing data. So the result in this table should be a stable evaluation of the performance. AUC is a balanced measurement for binary classification correctness combining both false positive and false negative. Its value ranges between 0 to 1, with 1 as the perfect classification. AUC and R^2 values demonstrate good regression results for all SCTG groups. Predicted network density is higher than actual network density, indicating a high false positive rate. This is desirable because a great portion of the predicted "existing" links will be estimated to be low flow value (thus filtered out by threshold) or abandon by the optimization process.

Table 4.6: Logistic regression performance

SCTG	average R^2	10-fold cross-validation AUC	predicted density	True density
1	0.86	0.93	0.20	0.077
2	0.82	0.91	0.23	0.076
3	0.694	0.78	0.42	0.255
4	0.76	0.85	0.34	0.195
5	0.76	0.84	0.39	0.26
6	0.75	0.84	0.40	0.27
7	0.77	0.85	0.51	0.51

4.3.2 Gamma regression

Gamma regression performance evaluation For Gamma regression, Drop-in-deviance test (*Ramsey and Schafer*, 2012) was used to assure every covariate in the model is significant. Null deviance test (*Ramsey and Schafer*, 2012) is conducted with 7 Gamma regressions and all have p-value = 0, showing the significance of the models. For saturated deviance tests (*Ramsey and Schafer*, 2012), however, the null hypothesis that current model is adequate compared to saturated model is rejected for all SCTG categories. This result means the current Gamma regression model has covariates that all informative for predicting the flow strength, but there is still missing information that could be included to further improve the Gamma regression. This is expected due to four reasons. First and most important, when estimating a flow from location i to j, existing flows related to i or j are unknown.

Imagine that even when i is a major soy producer and j has large soy processing plants, if i already send most of its production to other locations and j has already imported all needed soy from a third location, the soy flow from i to j is unlikely to be high. However, when we estimate the flow from i to j, we have no information about existing flows associated with them. Second, input data is truncated. Some commodity flow data existing between FAF levels is not available and treated as zero. Also flow data below a threshold is not collected or reported in the input data. Independent variables also have error and noise. For instance, exact transportation distance between FAF locations are not possible to be collected. Because FAF is a large area usually containing multiple counties and many transportation modes exist and even within the same mode there is no information about which commodity takes which route. Third, there are many variables impacting food flow and not all of them can be collected and considered. Even collected data is based on rough estimates found across multiple government agents. Forth, we have not consider nonlinear models. The first two reasons are the most important since we have also tested the data with neural network regression, the performance is not significantly improved. Neural network regression is considered a strong learner which can best fit the pattern existing in the data itself. We have tested with different depth of neural network regression until the performance stopped improving. Results show that it stops improving at around 2 layers, and no performance comparable to saturated model. This is an evidence that there is not enough information in the dataset itself to perfectly predict the flow strength, which is expected given its complexity and the two reasons listed. So the resulting gamma regression model is a significant model with all significant variables, but not perfect performing mainly due to the limitation of the existing data.

4.3.3 Sensitivity analysis

Comparable to any sensitivity analysis, the goal is to understand how the input variable impacts the output. In order to measure this impact, some metrics have been invented, among which first-order sensitivity and total-order sensitivity are popular choices. First order sensitivity, $S_j = V_j/V$, or "first-order sensitivity index" as named in Sobol's method, measures the contribution of input X alone to the output variance. That means no impact through interaction is considered. For example, in a system $y = 3x_1 + x_1 \times x_2$, first order sensitivity of x_1, only considers effect of $3x_1$ and $x_1 \times E(x_2)$, where x_2 averaged out. In comparison, total order sensitivity of x_1, effect of all the group of variables that contain x_1, would consider impact of both $3x_1$ and $x_1 \times x_2$.

As result in Table 4.7 demonstrates, for instance, SCTG3, its first order sensitivity is 0.046, meaning 4.6% of the variance of SCTG3 commodity flow is impacted by variable distance

when keeping all other variables as their average. As for its total order sensitivity 0.88, it means 88% of the output variance/output value change is related to distance.

Table 4.7: Sensitivity analysis result*

SCTG	Parameter	First	Total
1	pop_ori_2012	0.006469	0.270787
1	gdp_des_usd	0.000582	0.254088
1	dist	0.375566	0.90038
1	des_Seafood_product_preparation _and_packaging	0.000005	0.358891
1	ori_ANIMAL	0.006901	0.379664
1	des_Rendering_and_meat_byproduct _processing	0.00241	0.365603
1	gdp_ori_usd	0.000539	0.37031
1	ori_HOGS	0.007869	0.376055
1	des_Meat_processed_from_carcasses	0.00074	0.363384
1	des_Poultry_processing	0.001413	0.36244
2	pop_ori_2012	0.000068	0.5708
2	gdp_des_usd	0.000078	0.582319
2	prod_ori_ton	0.046206	0.643945
2	des_GOATS	0.000023	0.565999
2	des_Fruit_and_vegetable_canning	0.000312	0.571901
2	prod_des_ton	0.004944	0.600361
2	des_Other_animal_food _manufacturing	0.000014	0.581595
2	gdp_ori_usd	0.000176	0.551962
2	dist	0.260171	0.854361
2	des_CATTLE	0.018527	0.602311
3	pop_ori_2012	0.000009	0.857426
3	des_Seasoning_and_dressing _manufacturing	0.000002	0.863362
3	prod_ori_ton	0.000555	0.859017
3	des_Fruit_and_vegetable_canning	0.000050	0.868683
3	des_Frozen_food_manufacturing	0.000042	0.857577
3	prod_des_ton	0.00021	0.856795
3	dist	0.045991	0.880136
3	pop_des_2012	0.000261	0.858655
4	pop_ori_2012	0.014159	0.818108
4	des_CATTLE	0.001491	0.773181

Table 4.7 (cont)

4	prod_ori_ton	0.002351	0.763449
4	des_GOATS	0.000041	0.762689
4	prod_des_ton	0.000692	0.76785
4	des_Ice_cream_and_frozen_dessert _manufacturing	0.000032	0.758717
4	des_Other_animal_food _manufacturing	0.00007	0.754004
4	gdp_ori_usd	0.002827	0.787418
4	des_SHEEP	0.000041	0.757048
4	dist	0.070484	0.862593
4	pop_des_2012	0.000153	0.756847
5	ori_GOATS	0.000003	0.272272
5	ori_CHICKENS	0.001001	0.281093
5	dist	0.25944	0.895846
5	ori_Meat_and_meat_product _merchant_wholesalers	0.000001	0.390204
5	des_Food_and_beverage_stores	0.000001	0.402591
5	des_Wholesale_trade	0.000002	0.40281
5	ori_Animal_slaughtering_and _processing	0.000001	0.390964
5	ori_CATTLE	0.001766	0.416557
5	ori_Poultry_and_poultry_product _merchant_wholesalers	0.000001	0.392036
5	des_Other_animal_food _manufacturing	0.000153	0.407942
5	gdp_des_usd	0	0.386519
5	ori_Seafood_product_preparation _and_packaging	0	0.387662
5	ori_HOGS	0.001871	0.395909
5	prod	0.011013	0.431085
5	pop_des_2012	0.000001	0.403995
6	des_General_merchandise_stores	0.007318	0.569113
6	gdp_des_usd	0.000004	0.546176
6	ori_Grain_and_field_bean_merchant _wholesalers	0.001061	0.551941
6	des_Support_activities_for _transportation	0.00057	0.550576
6	dist	0.198914	0.892006
6	des_Snack_food_manufacturing	0.000007	0.585921

Table 4.7 (cont)

6	des_Sugar_and_confectionery_product _manufacturing	0.000138	0.57486
6	ori_Grain_and_oilseed_milling	0.000255	0.601096
6	ori_Bakeries_and_tortill a_manufacturing	0.000005	0.574844
6	gdp_ori_usd	0.000006	0.609778
6	des_Fats_and_oils_refining_and _blending	0.000076	0.585261
6	des_GOATS	0.000006	0.601542
6	prod	0.001202	0.583832
6	pop_des_2012	0.000001	0.578441
7	pop_ori_2012	0.000002	0.594573
7	des_General_merchandise_stores	0.000003	0.581167
7	dist	0.028465	0.87805
7	des_Dry_condensed,_and_evaporated _dairy_product_manufacturing	0.000025	0.723912
7	ori_CATTLE	0.000228	0.730746
7	ori_Soft_drink_and_ice_manufacturing	0.000022	0.722588
7	des_Other_animal_food _manufacturing	0.000021	0.720691
7	ori_Sugar_and_confectionery_product _manufacturing	0.000021	0.722398
7	des_Gasoline_stations	0.000432	0.736112
7	des_Fats_and_oils_refining_and _blending	0.000036	0.72287
7	ori_SHEEP	0.00003	0.722314
7	prod	0.000684	0.746406
7	ori_Other_food_manufacturing	0.00004	0.723231
7	des_Wholesale_trade	0.000044	0.721048
7	ori_TURKEYS	0.00004	0.722328
7	des_Coffee_and_tea_manufacturing	0.000035	0.723463
7	des_All_other_food_manufacturing	0.000043	0.721745
7	ori_CHICKENS	0.000043	0.72255
7	des_Food_and_beverage_stores	0.000033	0.721413
7	des_Soft_drink_and_ice_manufacturing	0.000052	0.724928
7	ori_Dairy_product_(except_dried_or _canned)_merchant_wholesalers	0.00004	0.721626
7	ori_Fresh_fruit_and_vegetable _merchant_wholesalers	0.000034	0.725305

Table 4.7 (cont)

7	ori_All_other_miscellaneous_chemical _product_and_preparation_manufacturing	0.000042	0.724323

* Sensitivity analysis for 1-7 SCTGs are listed. Parameter is the variable in the model. Notation on parameters can be found in table 4.3. *First* means first-order sensitivity, and *Total* means total-order sensitivity.

In the analysis result, we have the following findings 1. Across all SCTG categories, distance is the most influential variable to the output variance in terms of both first-order sensitivity and total order sensitivity. 2. Based on the criterion of total-effect index, all variable are important factors. This a not surprising because in our modeling process, we have used LASSO method to remove less important variables to avoid overfitting. 3. Some input variables have very small first-order effect index. This result indicates that some variables impact the result not by themselves but through interactions with other variables. Given that distance has such a large impact on the result, it strongly indicates that distance/convenience of transportation has a great impact on the transportation/trade volume. Potentially, it means improving current food supply chain system would result in great increase of trade/commodity transportation between counties for at least 1-7 SCTGs. As the result indicates, interactions between variables are responsible for a high percent of the output variance. However, an extension of this model to include interaction terms is beyond the scope of this paper. Because our goal is not to create the perfect model but to propose a framework, with which we can continue to improve. Also, with interactive terms, it becomes impossible to compare with gravity model of trade.

4.3.4 Flow estimation at county level results

There are 132 nodes in the FAF census data and 3,136 nodes in the county model results (see Table 4.8). So, we do not model 6 of the 3,142 counties in the United States due to lack of data for these counties (refer to the Supporting Information for list of these counties). There are 11,551 links in the FAF data out of a potential 17,292 links, leading to a density of 0.668.

Table 4.8: Network properties of food flows within the United States. Properties are provided for each SCTG group at the FAF and county spatial resolution. The mass flux of FAF scale self loops are included, as this mass is distributed amongst counties within those FAF zones. However, self loops are excluded from the # links and density network metrics, since self loops are not modeled at the county spatial scale.

FAF				
SCTG	# Nodes	# Links	Mass [kg]	Density
1	132	1,327	89.35E+9	0.077
2	132	1,314	789.16E+9	0.076
3	132	4,408	397.51E+9	0.255
4	132	3,364	258.60E+9	0.195
5	132	4,554	76.76E+9	0.263
6	132	4,635	101.42E+9	0.268
7	132	8,797	559.00E+9	0.509
Total	132	11,551	2.27E+12	0.668
County				
SCTG	# Nodes	# Links	Mass [kg]	Density
1	3,136	14,436	89.35E+9	0.001
2	3,136	19,622	789.16E+9	0.002
3	3,136	16,680	397.51E+9	0.002
4	3,136	18,567	258.60E+9	0.002
5	3,136	30,626	76.76E+9	0.003
6	3,136	30,779	101.42E+9	0.003
7	3,136	73,877	559.00E+9	0.008
Total	3,136	162,957	2.27E+12	0.091

Network density The Food Flow Model estimates 162,957 non-zero links at the county scale out of a potential 9,869,022 links. This means that the density of the county scale food flows is 0.091. The inferred network density at the county scale is much less than the empirical density at the FAF scale. However, the density at the county scale would likely be lower than the FAF scale, since this finer spatial resolution makes it unlikely that most counties would connect with one another directly and would instead transit through hubs. We also expect the county scale density to be smaller since self-loops are not modeled, but they are included in the FAF data. For example, the Remainder of Illinois reports a flow from itself as the origin to itself as the destination. Conversely, the Food Flow Model does not estimate a self-loop for Champaign County, Illinois. Of importance, note that the mass balances for each SCTG commodity class between the county and FAF spatial scales as required (see Table 4.8).

Fig 4.4 maps food inflows and outflows at the FAF and county spatial scales. The spatial trends compare reasonably well between FAF and county spatial scales. For example, note that California and the Great Lakes region are major outflow locations in the FAF data (see Fig 4.4A). Counties within these FAF areas are also locations of high food outflow in the nation (see Fig 4.4C). Similarly, the counties that are estimated to receive the most inflows of food correspond to the locations of FAF zones with high food receipts (compare Fig 4.4B with Fig 4.4D). This indicates that the Food Flow Model is maintaining the broad spatial trends observed at the FAF spatial scale as designed. Note that the mass transfer at each scale is different, which means that the scale on the color bar will also be different. The masses being transferred at the county scale are smaller than at the FAF level, because the mass at the FAF level has been distributed to counties, and we do not expect a single county within a FAF zone to transfer the entirety of the food mass. However, we are now able to infer how food flows are distributed across counties, which we are not able to observe at the FAF scale.

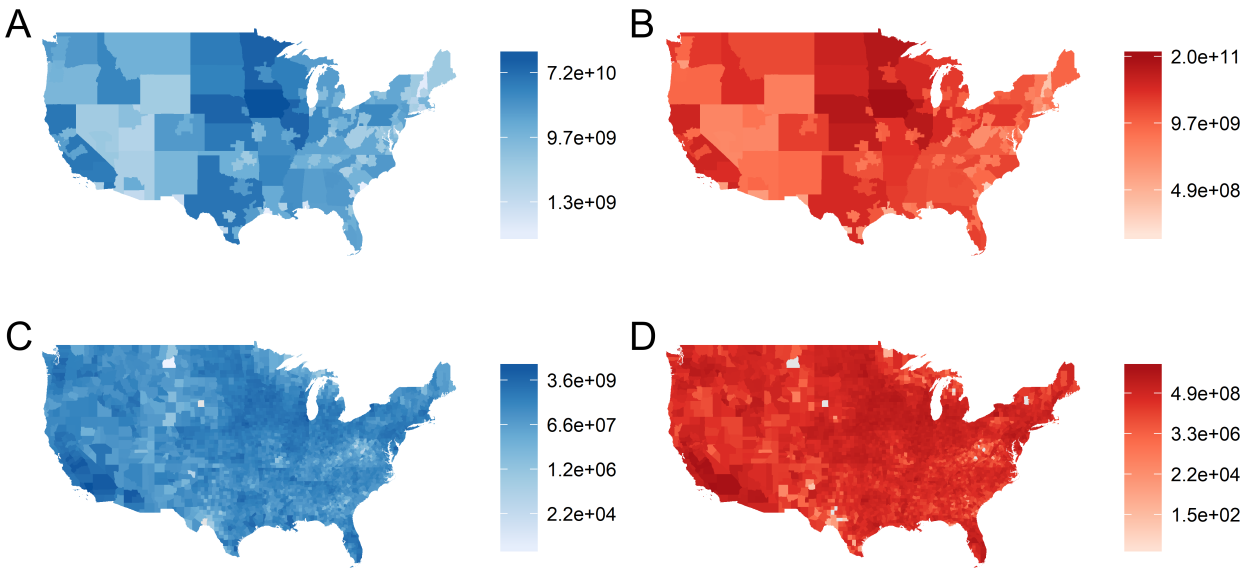

Figure 4.4: Maps of food outflows and inflows within the United States. Maps depict food (A) outflows at the FAF scale, (B) inflows at the FAF scale, (C) outflows at the county scale, and (D) inflows at the county scale. These maps depict aggregate food flows [tons]. Maps for specific food commodity groups are provided in the Supporting Information document.

Table 4.9 ranks the top outflow and inflow locations by spatial scale. The county scale estimates are broadly consistent with the FAF scale census data. Our model estimates that several California counties are the largest in terms of outflows and inflows. For example, Los Angeles county is predicted to be the largest origin and destination node at the county scale, despite the fact that the Remainder of Iowa FAF zone is the largest origin and destination node at the FAF scale (refer to Table 4.9). This indicates that the large Remainder of Iowa link is more evenly distributed amongst the counties within Iowa, while the mass flux within the state of California is distributed in a fairly heterogeneous manner amongst its counties. This is likely due to the high heterogeneity in production and consumption hubs within California. This may also be a function of the linear programming algorithm that minimizes travel cost. Distances between a northern California county and southern California county are larger than distances across the state of Iowa, for example. This means that shipping food goods from norther to southern California will face relatively high transportation costs. This will likely force more aggregated local flows in California since the objective function will be heavily penalized for shipping across the large state. Additionally, counties in the western portion of the United States, such as California, are larger than counties in the east, which will also lead to more aggregation at the county spatial scale.

Table 4.9: Ranking of food flows within the United States by mass [kg]. The top 10 food outflow and inflow FAF zones and counties are provided. Note that aggregate food flows are provided. The ranking for specific food commodity group is provided in the Supporting Information.

FAF				
Rank	Outflow	Mass [kg]	Inflow	Mass [kg]
1	Remainder of Iowa	6.47E+10	Chicago-Naperville, IL-IN-WI CFS Area (IL Part)	3.58E+10
2	Remainder of Minnesota	4.90E+10	New Orleans-Metairie-Hammond, LA-MS CFS Area (LA Part)	3.37E+10
3	Remainder of Kansas	4.82E+10	Remainder of California	3.36E+10
4	Remainder of Illinois	4.63E+10	Remainder of Illinois	2.92E+10
5	Remainder of Nebraska	4.28E+10	Remainder of Minnesota	2.89E+10
6	Remainder of California	3.45E+10	Remainder of Texas	2.85E+10
7	Remainder of Indiana	2.94E+10	Remainder of Iowa	2.84E+10
8	Chicago-Naperville, IL-IN-WI CFS Area (IL Part)	2.52E+10	Minneapolis-St. Paul, MN-WI CFS Area (MN Part)	2.59E+10
9	Remainder of Wisconsin	2.37E+10	Omaha-Council Bluffs-Fremont, NE-IA CFS Area (NE Part)	2.50E+10
10	Remainder of North Dakota	2.26E+10	Los Angeles-Long Beach, CA CFS Area	2.39E+10
County				
Rank	Outflow	Mass [kg]	Inflow	Mass [kg]
1	Los Angeles County, CA	1.49E+10	Los Angeles County, CA	2.08E+10
2	Fresno County, CA	1.14E+10	Orange County, CA	1.25E+10
3	Riverside County, CA	1.12E+10	Fresno County, CA	1.19E+10
4	San Bernardino County, CA	1.12E+10	Maricopa County, AZ	1.03E+10
5	San Joaquin County, CA	9.88E+09	San Bernardino County, CA	9.95E+09
6	Tulare County, CA	9.74E+09	Stanislaus County, CA	9.36E+09
7	Merced County, CA	8.78E+09	Cook County, IL	9.01E+09
8	Stanislaus County, CA	8.58E+09	Douglas County, NE	8.75E+09
9	Cattaraugus County, NY	7.12E+09	Madera County, CA	7.06E+09
10	Richland County, ND	6.26E+09	Niagara County, NY	6.82E+09

Fig 4.5 maps food flows at the FAF and county spatial scales. Links are shown for all FAF flows (11,551 existing links) and for the largest 0.1% of county estimates (162,957 existing links). These maps depict aggregate food flows and the points are graduated based on inflows. The general spatial trends between the FAF and county spatial scales compare reasonably well in Fig 4.5. For example, note that the strong connectivity between the corn/soy belt and the port of New Orleans exists in both the FAF data (see Fig 4.5A) and the county modeled results (see Fig 4.5B). Similarly, the links between the New York area and the Great Lakes, as well as the connections from the grain belt to California, are shown in both Fig 4.5A and B. The density is much higher for the FAF data than inferred county results (refer to Table 4.8). However, this is sensible, since the spatial scale is so much larger (by definition) in FAF, there will be more connectivity. The county flow results were additionally pruned to exclude links with fluxes <1kg, further reducing the estimated density at this scale.

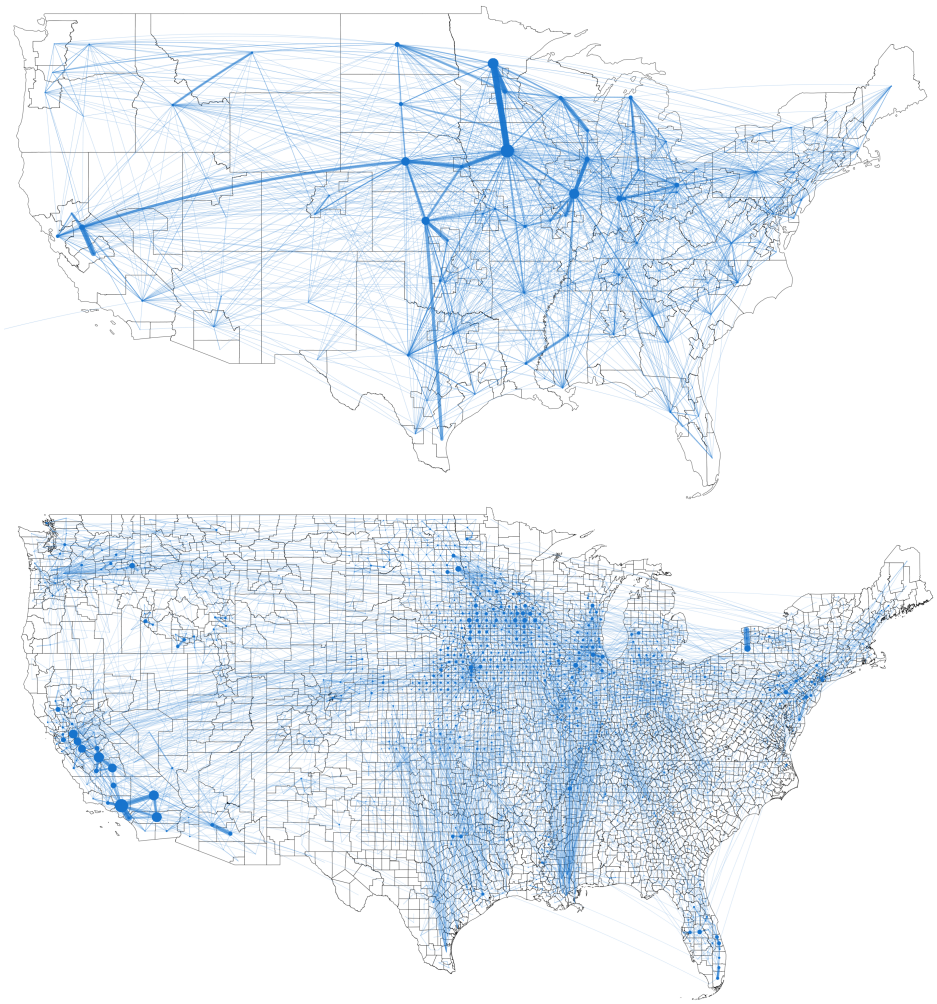

Figure 4.5: Maps of food flows within the United States. Maps depict the food flow network at the (A) FAF and (B) county scale. Points are graduated based on inflows. Links are shown for all FAF data (11,551 total) and for 0.1% of county estimates. These maps depict the topology of aggregate food flows [tons].

The model estimates that most of the largest county links are located within close proximity to one another (see Table 4.10). This means that the largest mass fluxes at the county scale are largely driven by distance. Again, the model estimates the mass fluxes within California are some of the largest links in the county. For example, the transfer of food from Los Angeles County, CA to Orange County, CA is the largest link at the county scale, while Orange County, CA to Los Angeles County, CA is the 8th largest link. In fact, half of the 10 largest links are estimated to be within California. This model results is sensible due to the large mass fluxes reported in the FAF data combined with the large spatial heterogeneity in production and attraction factors for food. Even though other FAF zones tend to transfer larger masses than those within California, the counties within those FAF zones are

more homogeneous. For example, one of the largest FAF transfers is Remainder of Illinois to Chicago-Naperville. Due to relatively limited diversity in crop type and production patterns across Illinois, it is unsurprising that county level food transfers from the Remainder of Illinois to Chicago-Naperville exhibit a similar level of homogeneity. This homogeneous production and distribution would inhibit a handful of counties from supplying Chicago-Naperville all of its food goods from within the state. In this way, the more heterogeneous distribution to counties within California leads them to have the highest ranking at the link level.

Table 4.10: Ranking of food flows within the United States by mass [kg]. The top 10 links at the FAF and county scales are provided. Note that aggregate food flows are provided. The ranking for specific food commodity group is provided in the Supporting Information.

FAF		
Rank	Link	Mass [kg]
1	Remainder of Minnesota → Remainder of Iowa	1.32E+10
2	Remainder of Wisconsin → Milwaukee-Racine-Waukesha, WI CFS Area	1.30E+10
3	Remainder of Kansas → Corpus Christi-Kingsville-Alice, TX CFS Area	1.29E+10
4	Remainder of Iowa → Remainder of Minnesota	1.26E+10
5	Minneapolis-St. Paul, MN-WI CFS Area (MN Part) → Remainder of Minnesota	1.19E+10
6	Remainder of Nebraska → Remainder of California	1.10E+10
7	Remainder of Minnesota → Minneapolis-St. Paul, MN-WI CFS Area (MN Part)	1.01E+10
8	Remainder of California → Fresno-Madera, CA CFS Area	1.04E+10
9	Remainder of Louisiana → Remainder of Mississippi	1.02E+10
10	Remainder of Illinois → Chicago-Naperville, IL-IN-WI CFS Area (IL Part)	9.07E+09
County		
Rank	Link	Mass [kg]
1	Los Angeles County, CA → Orange County, CA	6.11E+09
2	Cattaraugus County, NY → Niagara County, NY	5.60E+09
3	San Bernardino County, CA → Los Angeles County, CA	4.46E+09
4	Pinal County, AZ → Maricopa County, AZ	3.74E+09
5	San Joaquin County, CA → Stanislaus County, CA	3.18E+09
6	Merced County, CA → Stanislaus County, CA	2.96E+09
7	Camden County, NJ → Sussex County, DE	2.72E+09
8	Fresno County, CA → Madera County, CA	2.69E+09
9	Riverside County, CA → San Bernardino County, CA	2.68E+09
10	Los Angeles County, CA → Ventura County, CA	2.62E+09

Properties of generated county level network As shown in 4.11, in the estimated county network, degree distributions across 1 to 7 SCTG groups fit generalized exponential distribution and strength distributions also fit gamma distribution. This is surprising result because of x reasons. First, by applying logistic regression, there is no guarantee that the degree in the outcome network would follow a certain distribution. Further more, the flow strength estimation and distance optimization will potentially remove extra links. The fact that degree distribution in the outcome is indeed generalized exponential confirms that the

Gamma hurdle model has captured the pattern in data well, which validates our underlying Bernoulli trial assumption between a pair of locations, $P(link_{ij}|X_{ij}) = f(X_{ij})$ and our flow strength Poisson process assumption. Second, even we have proved that a food flow as a Poisson process between a location pair will lead to $flow_{ij} \sim Gamma(K, \theta)$ in 4.2.3, the extra mass balance constraint, and distance optimization in the simulation process makes it unnecessarily true. The fact that it is indicates the robustness of our assumption.

Table 4.11: Goodness of fit for flow degree and flow strength

SCTG	EP flow fit ajd R2	EP degree fit adj R2
1	0.98	0.99
2	0.99	0.98
3	0.96	0.99
4	0.97	0.99
5	0.99	0.99
6	0.97	0.99
7	0.97	0.99

Figure 4.6 shows detailed distributions.

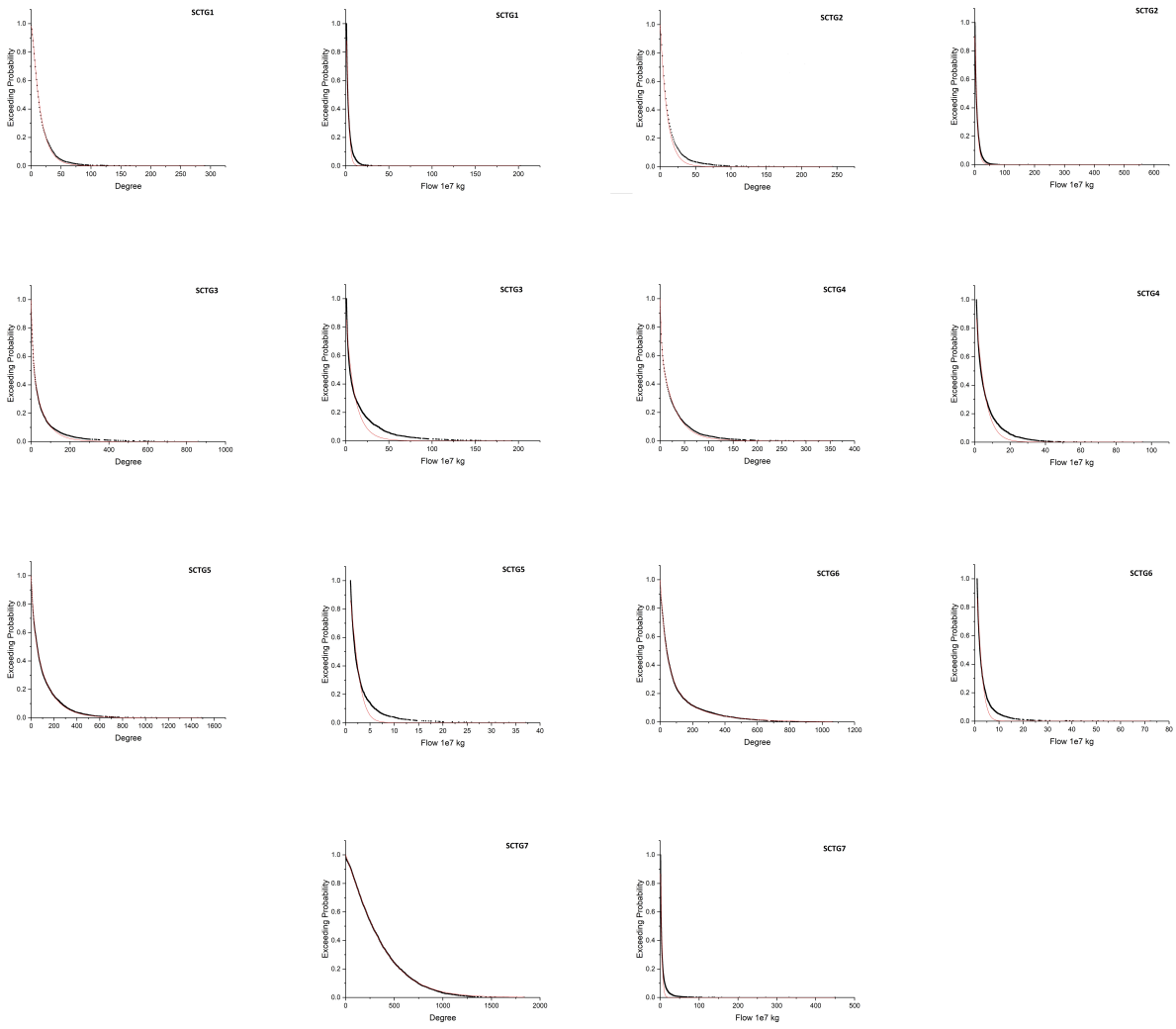

Figure 4.6: Degree distributions and strength distributions of SCTG 1 to 7 in estimated county network

From figure 4.7, we see right skewed degree distribution, strength distribution and betweenness distribution, just like what we have seen in the other three scales. Highest betweenness in county network is close to national and lower than village, indicating it is not as close to spoke network as village network does. Medium of clustering coefficient in estimated county network is close to it in village household network.

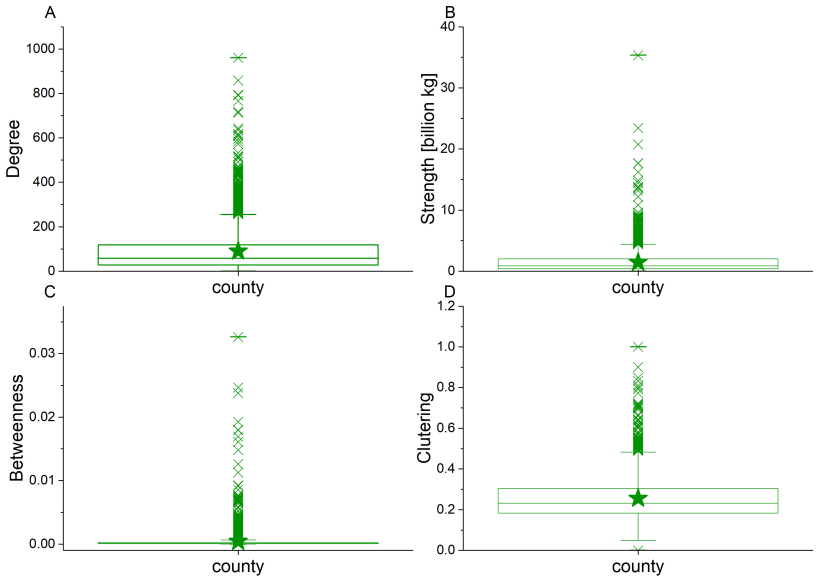

Figure 4.7: Estimated county commodity flow network boxplot

In figure 4.8, it is shows the relationship between degree and betweenness, $B = a \times k^b$, is preserved in estimated county commodity flow network with Adj $R^2 = 0.93$ for undirected network and Adj $R^2 = 0.94$ for directed network. While there is no imposition in our model to enforce this relationship, it emerges naturally from our assumption and regression model. This result is revealing a potential relationship between our assumptions and betweenness attribute and further confirms our theory: the link and flow formation process we proposed generates the network we have observed across scales.

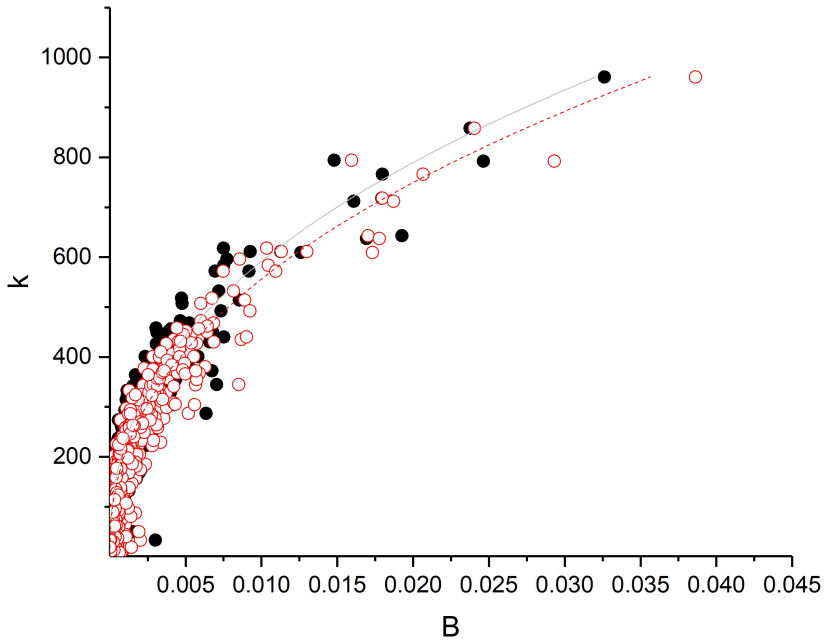

Figure 4.8: Degree vs betweenness in estimated county commodity flow network

4.3.5 Comparison with literature

We compare our results with the county-scale corn flows modeled by *Smith et al.* (2017). To our knowledge, this is the only other information on county-scale food flows in the United States. *Smith et al.* (2017) use a transportation optimization model to estimate corn flows between U.S. counties. To compare our results, we transform our estimates of SCTG 2 to estimates of corn by multiplying the SCTG 02 flows by the fraction of corn grains produced in each origin county as compared to total grain production. Grains here include the following crops: barley, buckwheat, corn (grain), corn (silage), millet (proso), oats, rice, rye, sorghum (grain), sorghum (silage), triticale, wheat, and wild rice. We compare results for Corn Belt states (Illinois, Indiana, Iowa, Kansas, Missouri, and Nebraska) to avoid inaccuracies in states which produce large quantities of other grains. Fig 4.9 maps total inflows and outflows for each Corn Belt county. Note that both inflow and outflow maps share a common scale. The spatial trends compare remarkably well between the two models.

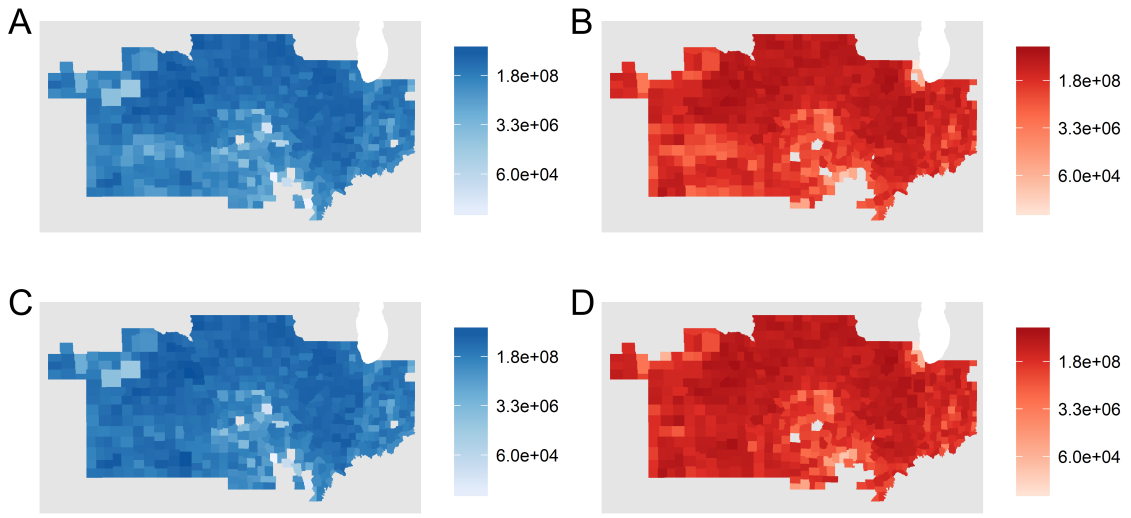

Figure 4.9: Comparison of maize flow [tons] maps for the Corn Belt of the United States. Maps show (A) maize inflows for our model, (B) maize outflows for our model, (C) maize inflows from *Smith et al.* (2017), and (D) maize outflows from *Smith et al.* (2017).

Fig 4.10 shows how our corn flows compare to the flows estimated by *Smith et al.* (2017). Note that maps in Fig 4.10 share the same scale and indicate that our model has more links and larger outflows for many counties. Our model has more links with smaller values. In particular, our model shows more links in Nebraska than does *Smith et al.* (2017), which is more concentrated in Illinois and Iowa.

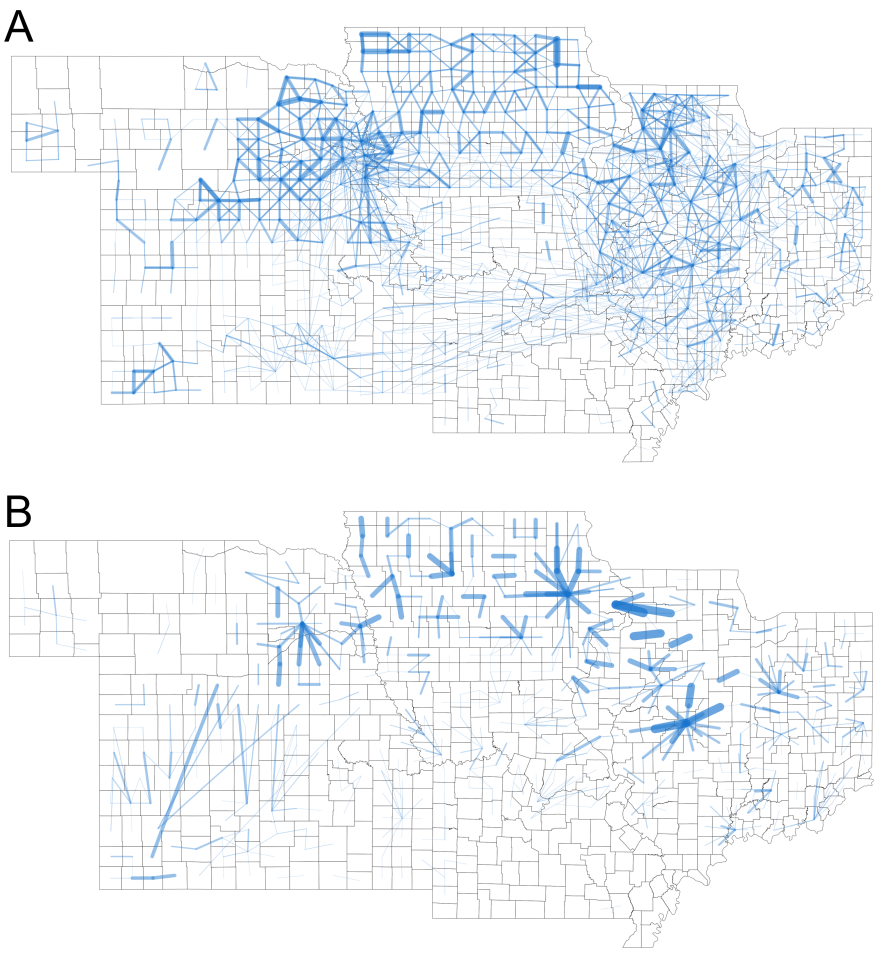

Figure 4.10: Comparison of maize flow networks [tons] for the Corn Belt of the United States. (A) The Food Flow Model for SCTG 02 scaled by maize production within each county. (B) Maize flows from *Smith et al.* (2017). Links represent maize flows between counties.

Table 4.12 provides Corn Belt state outflows, inflows, and intra-flows (flows from a state to itself). This information is provided for FAF data, *Smith et al.* (2017), our model without corn normalization, and our model with corn normalization. The total flows estimated differ between our model of corn and the *Smith et al.* (2017) model. It is important to note that our model without corn normalization replicates the raw FAF data (as designed). So, our model characterizes similar spatial trends as *Smith et al.* (2017), with the additional advantage of being constrained by FAF data. Importantly, *Smith et al.* (2017) use their estimated corn flows to perform a spatially explicit environmental impact analysis of the US corn supply chain. This application of lifecycle assessment to spatially refined corn flows highlights potential applications of similar methods to the spatially refined estimates of all food flows provided in this paper.

Table 4.12: Comparison of maize flows for the Corn Belt in the United States. Note that values for this study (without corn production scaling) are equivalent to the data reported by the Freight Analysis Framework (FAF).

This study (with corn production scaling)				
State	Outflows	Inflows	Intraflows	All Flows
IL	2.24E+09	4.23E+09	6.54E+10	7.19E+10
IN	2.20E+09	5.49E+08	2.24E+10	2.52E+10
IA	6.75E+09	2.31E+09	7.29E+10	8.19E+10
KS	4.40E+09	3.71E+09	1.98E+10	2.79E+10
MO	1.43E+09	2.94E+09	9.48E+09	1.39E+10
NE	5.30E+09	8.58E+09	6.48E+10	7.87E+10
This study (without corn production scaling)				
State	Outflows	Inflows	Intraflows	All Flows
IL	2.57E+09	5.17E+09	7.09E+10	7.87E+10
IN	2.28E+09	6.25E+08	2.32E+10	2.61E+10
IA	6.76E+09	2.32E+09	7.31E+10	8.21E+10
KS	6.53E+09	4.12E+09	4.47E+10	5.54E+10
MO	1.68E+09	3.56E+09	1.30E+10	1.83E+10
NE	5.64E+09	9.66E+09	6.73E+10	8.26E+10
Smith et al (2017) model				
State	Outflows	Inflows	Intraflows	All Flows
IL	6.97E+09	6.89E+07	2.31E+10	3.01E+10
IN	6.07E+07	1.29E+09	1.46E+10	1.59E+10
IA	7.90E+08	5.61E+09	4.84E+10	5.48E+10
KS	4.04E+08	1.92E+09	8.03E+09	1.04E+10
MO	3.29E+08	8.92E+08	3.99E+09	5.22E+09
NE	1.95E+09	7.12E+08	2.46E+10	2.73E+10
FAF				
State	Outflows	Inflows	Intraflows	All Flows
IL	2.57E+09	5.17E+09	7.09E+10	7.87E+10
IN	2.28E+09	6.25E+08	2.32E+10	2.61E+10
IA	6.76E+09	2.32E+09	7.31E+10	8.21E+10
KS	6.53E+09	4.12E+09	4.47E+10	5.54E+10
MO	1.68E+09	3.56E+09	1.30E+10	1.83E+10
NE	5.64E+09	9.66E+09	6.73E+10	8.26E+10

4.4 Conclusion

Based on consistent patterns we have observed in Chapter 3, we proposed our assumptions about the underlying processes in commodity flow network. Mathematical relationship between our assumptions and the patterns in degree distributions and strength distributions has between demonstrated. These assumptions are suggesting a statistical relationship between environmental variables and commodity flow existence and strength. Based on this relationship, we prosed Gamma hurdle model. We trained the model with data available at FAF level to quantitatively describe the relationship between environmental variables and commodity flow existence and strength. By verifying the model performance, we validate our assumption. Utilizing this model, we estimated the flow existence and strength at county level. By demonstrating the consistent patterns (degree distribution, strength distribution, betweenness vs degree relationship, and alignments of the results in county level with national and village level), we have further prove the relationship between our assumption and the network properties. That is, the processes illustrated by our mathematical assumption,

is responsible for the patterns we have observed across scales. The county level commodity flow estimates also have important scientific value to the academia, as the need for high resolution commodity flow data is essential in many researches.
Furthermore, to overcome the limitation of of no validation data at county level, we have compared our SCTG2 commodity flow estimate to the estimate from (*Smith et al.*, 2017) with distinctively different method, it shows remarkable similarity.

In gravity model of trade (*Burger et al.*, 2009), variables are limited to economic variables while our model has included variables related to food processing and manufacturing. These variables are selected from known factory products/procedures utilizing corresponding food commodities in their production. Thus, our model has more realistic relevance. We also observe that in the gravity model of trade, distance has a coefficient between -0.9 to -0.4, while in our model, the minimum distance coefficient, -1.35, is from SCTG 7. SCTG 7 is representing prepared foodstuff including fats and oils, such as milk, cheese and butter. This is indicating the transportation of these commodities is more impacted by distance.

Our model is trained for FAF zone flow data and then extrapolated (with other procedures) to locations without flow information (counties). This increases our confidence that our model is producing reasonable results, despite the fact that we do not have data available to validate against. We use high-resolution production data to enhance the realism of inferred results. Additionally, we force the flows of the counties that are located within each FAF zone to sum to the reported FAF flows. This constraint ensures that our county scale estimates are in accordance with FAF empirical information and lend additional reliability to Food Flow Model results. Nonetheless, for locations that have empirical information on food flows available, these would be preferred to our inferred values.

Future work could improve the realism of our algorithm. For example, more realistic distance matrices could be used, such as those that are constrained by available infrastructure. For example, future research could utilize distances between counties based upon the roadway network, rather than shortest paths. In fact, future research could take advantage of the mode information provided by FAF to further resolve these food flow estimates to specific infrastructure networks (i.e. road, rail, waterway). If this is accomplished, an inter-connected network model could be developed to reveal the specific infrastructure that is critical to the national food supply chain. Additionally, future research could combine these detailed food flow estimates with high-resolution footprint estimates to evaluate the water, carbon, and nutrient footprint of the national food supply chain in the United States.

Understanding variable interactions would be the next step to improve this model and would be put into the further work. Further, including interactive terms would make it impossible to compare our result to gravity models in economics. It would be interesting to

study the interactions between all variables as it indicates in the analysis result. However, this paper's focus is to create a method of modeling the transportation flow between all counties in USA. Understanding variable interactions would be the next step to improve this model and would be put into the further work. Further, including interactive terms would make it impossible to compare our result to gravity models in economics.

CHAPTER 5

CONCLUSION

5.1 Concluding remarks

This dissertation has striven to advance our understanding of food flow networks. Step by step, it starts from empirical description by characterizing a benchmark food flow network with a country. Then it expands the scope to multiple spatial scales where both invariant and variant patterns in food flow network emerge. Lastly, underlying statistical processes responsible for the consistent patterns are proposed, described with statistical distributions, quantified through statistical regression, and validated by the fact that the estimated network from these processes and their resulting model have preserved all the consistent patterns discovered in the scaling study.

In the empirical description step, we have explored the food flow network connecting sub-national locations across the United States. The US is the greatest food producer in the world, has comprehensive and open data, transportation infrastructure radiating in all directions, and no political or tariff interference, making it an ideal benchmark case study. The food flow networks of the US exhibits signatures of a social rather than technological networks, exemplified through triad significance profiling. At the same time, some structural "hubs" have been identified in this food flow network which are responsible for connecting places and routing food. Importantly, food flows within the US can be thought of as the most equal that food flows can expect to be, i.e. a "null model" for trade equality. This is because there is a similar culture, identical currency, and no trade barriers throughout the country, as opposed to the existence of these objects at the international trade scale. Equality of the US food flow network has been illustrated through Gini coefficient analysis and perfect equality does not exist, which makes sense, given that spatial dislocation in production and consumption exist even within a country, necessitating a heterogenous food flow system. This highlights that we should not expect the global trade system to exhibit perfect equality either.

While the case study of a food flow network within the US demonstrate the existence of some interesting properties in a specific food flow network, it is not generalizable. Hence,

we proceed to conduct a scaling analysis of food flow networks. To do this, we examine multiple spatial scales, including households in villages of Alaska, sub-national areas in the US, and countries in the world. Some common patterns in food flow networks have been discovered. The relationship between node connectivity and mass flux follows a power law across scales. Mean node connectivity and mass flux increase with increasing scale. A core group of nodes exists at all scales, but node centrality increases as the spatial scale decreases, indicating that some households are more critical to village food exchanges than countries are to global trade. The network structures of food flow systems provide a signature of their vulnerability and resiliency to disturbance. Extensive research has explored the implications of certain network structures for vulnerability and resiliency. For example, networks with a power law node degree distribution have been shown to be vulnerable to targeted attack, but resilient to random attack. Remarkably, the statistical distribution of node connectivity, as well as the distribution of mass flux are consistent across scales. Statistical distributions are generated from random processes. Thus the consistency of statistical distributions indicate the underlying random processes responsible for these distributions may be the same.

We build on the knowledge of statistical network properties gained in the case study and scaling study to develop a new model of food flows. Specifically, we estimate food flows between all county pairs within the US. To do this, we develop the Food Flow Model, a data-driven methodology to estimate spatially explicit food flows. The Food Flow Model integrates machine learning, network properties, production and consumption statistics, mass balance constraints, and linear programming. Specifically, we downscale empirical information on food flows between 132 Freight Analysis Framework (FAF) locations (17,292 potential links) to the 3,142 counties and county-equivalents of the United States (9,869,022 potential links).

We use logistic regression for link prediction in the Food Flow Model. Thus, effectively, the Bernoulli process is assumed to be the random process that is generating the degree distribution of food flow networks. This Bernoulli process has a success probability decided by environmental variables between the two places, such as distance, GDP, and crop production, and varies across different pairs, which can be described as $P(link_{ij}|X_{ij}) = f(X_{ij})$ with $f(X_{ij})$ denoting the relationship between variables and success probability. Through logistic regression, variables have been selected and corresponding coefficients in $f(X_{ij})$ have been estimated. As flow strength $flow_{ij}$ follows the Gamma distribution, empirically shown in 4.2.3, the mass flux process can be modeled as a Poisson process, $Pois(\lambda)$, where as each unit of food flows from origin to destination, it has a certain probability that this unit of food is "effective" towards fulfilling the requirements between these two locations. Aggregating all units of food together between two places, we get a conditional Gamma distribution

between them, $P(flow_{ij}|X_{ij}) = Gamma(k(X_{ij}), \theta)$. That is, the environmental variables between two places decides the quantity ($k(X_{ij})$) of successful flow required. And food flows from origin to destination with a certain success rate, until the required amount of successful flow happens. Through Gamma regression, we estimated different $k(X_{ij})$ functions for different SCTGs, and θ for the whole network.

By combining these two model, we obtain a Gamma hurdle model. We apply this gamma hurdle model to estimate the county level food flow network. We have shown that, remarkably, our model preserves the degree distribution, strength distribution, as well as the relationship between degree and strength. This outcome provides further evidence that our assumption about the underlying random processes responsible for food flow network generation are well founded. The goal of our model was to estimate food flows in locations without data, which provides us with the challenge of no data available for validation. However, we have compared our estimates with *Smith et al.* (2017) for corn. Our model compares well with *Smith et al.* (2017) despite differences in underlying modeling approaches, providing increased evidences for the believability of our model approach.

This has significant implications. The Food Flow Model is the first model to explain both the network topology and mass flux based upon a consistent handling of the underlying random processes. Through the Food Flow Model, we have connected the micro-scale attributes of specific locations with the macro-scale emergent structure of food flow networks. This model can be applied to areas where food flow data is missing. Such food flow estimates can be used in future work to improve our understanding of vulnerabilities within a national food supply chain, determine critical infrastructures, and enable spatially detailed footprint assessments.

5.2 Suggestions for future research

There are many opportunities for future research to improve upon the work presented in this dissertation. First, we hope that future work will explore the validity of applying models with parameters estimated at a larger spatial resolution (i.e. FAF) to a smaller spatial resolution (i.e. county level) within a country. It is reasonable to assume that the relationship between $f(X_{ij})$ and $k(X_{ij})$ will be consistent between spatial resolutions. However, it is impossible to verify this for the United States, wince county level food flow data is not available. This relationship could be confirmed in places that have evidence across multiple spatial resolution. Additionally, future work could collect and integrate food consumption data into the model. If both production and consumption data were to be available, then an additional mass balance restriction could be incorporated into the optimization to further improve the result. Another major improvement would be to change

the generalized linear model into a generalized nonlinear model in the gamma hurdle model. For logistic regression, the link function can be improved by replacing the linear model with piecewise nonlinear model, to consider independent factors in each location's contribution to the linkage.

Another improvement to the Food Flow Model would be to incorporate degree into the process for estimating network topology. This could be done through Expectation Maximization (EM) algorithm. The output of the current method could be used as the input to an EM algorithm. The EM algorithm would use node degree in this output as an additional input to the existing model and estimate a new food flow network. This process would be repeated until the change in model input and output in the EM algorithm is negligible. This should further improve the accuracy of the model result. More studies about the network topology can be conducted with generalized exponential distribution. Simulations of the stochastic processes can be used to generate a theoretical network topology. With this theoretical network, we can compare the clustering coefficients as well as degree distribution against the real-world food flow networks. If the theoretical network preserves the properties of network topology exhibited by the empirical network, this mechanism can be interpreted as a new explanation for network topology formation in food flow networks.

We provide estimate of county-scale food flows for the year 2012. This was an exceptional drought year in the United States. Importantly, the drought impacts should be captured by the input data within our model already, meaning our model was able to incorporate these notable conditions. However, it is possible that the regression models will be specific to each time period. Our modeling framework is general and would apply in other years; however, the Food Flow Model should be run in each new time period to ensure the most accurate results. In fact, we suggest that comparing model structure and performance in a different time period (i.e. non-drought year) is an an important area of future research.

Future work could incorporate more realistic distance matrices to the Food Flow Model, such as those that are constrained by available infrastructure. For example, future research could utilize distances between counties based upon the roadway network, rather than shortest paths. In fact, future research could take advantage of the mode information provided by government databases to further resolve these food flow estimates to specific infrastructure networks (i.e. road, rail, waterway). If this is accomplished, an inter-connected network model could be developed to reveal potential vulnerabilities and resiliences of the national food supply chain. Additionally, future research could combine these detailed food flow estimates with high-resolution footprint estimates to evaluate the water, carbon, and nutrient footprint of the national food supply chain in the United States.

APPENDIX A

SUPPLEMENTARY MATERIALS

A.1 Figures and Tables for Chapter 2

A.1.1 Supplementary Methods: Network Statistics

In this section we provide the equations that we use to calculate node nearest neighbor degree, and clustering and perform the network triad analysis. Average nearest neighbor degree (knn) measures the affinity of a node to connect to high- or low-degree neighbors, or the network correlation structure (*Watts*, 1999; *Jackson*, 2008). When direction is taken into account, weighted values of knn can be measured with four directional pairs: in-in (ii), out-out (oo), in-out (io), and out-in (oi). Following (*Konar et al.*, 2011), we use the following weighted, directed knn definitions:

$$knn_i^{W(ii)} = \frac{1}{s_{in_i}} \sum_{j \in V_{in_{(i)}}} w_{ji} k_{in_j} \tag{A.1}$$

$$knn_i^{W(oo)} = \frac{1}{s_{out_i}} \sum_{j \in V_{out_{(i)}}} w_{ij} k_{out_j} \tag{A.2}$$

$$knn_i^{W(io)} = \frac{1}{s_{in_i}} \sum_{j \in V_{in_{(i)}}} w_{ji} k_{out_j} \tag{A.3}$$

$$knn_i^{W(oi)} = \frac{1}{s_{out_i}} \sum_{j \in V_{out_{(i)}}} w_{ij} k_{in_j} \tag{A.4}$$

where $j \in V_(i)$ indicates the j neighbors of node i for a given trade direction. For example, $\in V_{in_{(i)}}$ indicates the neighbors from which node i imports.

The clustering coefficient measures the degree to which nodes tend to cluster together or form closed triangles (*Watts*, 1999). When direction is taken into account, there are eight possible combinations of the local clustering coefficient that fall into four categories

(see (*Fagiolo*, 2007) for a complete description and representation): C_{in}, C_{out}, C_{cyc}, and C_{mid}. Following (*Konar et al.*, 2011), we use the following weighted, directed definition for the local clustering coefficient:

$$C^W_{in_i} = \sum_{j,h \in V_{in_{(i)}}} \frac{(w_{ji} + w_{hi}) a_{jh|hj}}{2 \cdot s_{in_i}(k_{in_i} - 1)} \tag{A.5}$$

$$C^W_{out_i} = \sum_{j,h \in V_{out_{(i)}}} \frac{(w_{ij} + w_{ih}) a_{jh|hj}}{2 \cdot s_{out_i}(k_{out_i} - 1)} \tag{A.6}$$

$$C^W_{cyc_i} = \sum_{j \in V_{out_{(i)}}} \sum_{h \in V_{in_{(i)}}} \frac{(w_{ij} + w_{hi}) a_{jh}}{s_{tot_i}(k_{tot_i} - 1)} \tag{A.7}$$

$$C^W_{mid_i} = \sum_{j \in V_{in_{(i)}}} \sum_{h \in V_{out_{(i)}}} \frac{(w_{ih} + w_{ji}) a_{jh}}{s_{tot_i}(k_{tot_i} - 1)} \tag{A.8}$$

where $a_{jh|hj}$ indicates that a closed triangle is formed if a link exists $j \to h$ or $h \to j$. For this reason, C_{in} and C_{out} are divided by two in the above equations to avoid double counting closed triangles. We define $k_{tot} = k_{in} + k_{out}$ and $s_{tot} = s_{in} + s_{out}$.

Triads are three-node directed sub-graphs. Triad frequencies of empirical networks are compared to frequencies in a random network to arrive at a z-score for each triad type:

$$z = \frac{N^{actual} - N^{random}}{std(N^{random})} \tag{A.9}$$

In order to normalize z-scores and ensure that they $\in$ [0,1], we follow (*Milo et al.*, 2004) and apply the following equation:

$$SP_i = \frac{z_i}{(\sum z_i^2)^{\frac{1}{2}}} \tag{A.10}$$

The normalized z-score is plotted for each triad type to obtain the triad significance profile (TSP) of the network. TSPs can be compared across networks, since z-scores are normalized and dimensionless.

A.1.2 Supplementary Tables

Table A.1: List of 123 CFS Areas (i.e. nodes of the network) in alphabetical order.

index	CFS Area
1	Alaska
2	Albany-Schenectady-Amsterdam
3	Arkansas
4	Atlanta-Sandy Springs-Gainesville
5	Austin-Round Rock
6	Baltimore-Towson
7	Baton Rouge-Pierre Part
8	Beaumont-Port Arthur
9	Birmingham-Hoover-Cullman
10	Boston-Worcester-Manchester, MA part
11	Boston-Worcester-Manchester, RI part
12	Buffalo-Niagara-Cattaraugus
13	Charleston-North Charleston
14	Charlotte-Gastonia-Salisbury
15	Chicago-Naperville-Michigan City, IL part
16	Chicago-Naperville-Michigan City, IN part
17	Cincinnati-Middletown-Wilmington
18	Cleveland-Akron-Elyria
19	Columbus-Marion-Chillicothe
20	Corpus Christi-Kingsville
21	Dallas-Fort Worth
22	Dayton-Springfield-Greenville
23	Delaware
24	Denver-Aurora-Boulder
25	Detroit-Warren-Flint
26	El Paso
27	Grand Rapids-Muskegon-Holland
28	Greensboro-Winston Salem-High Point
29	Greenville-Spartanburg-Anderson
30	Hartford-West Hartford-Willimantic
31	Honolulu
32	Houston-Baytown-Huntsville

Table A.1 – continued from previous page

index	CFS Area
33	Idaho
34	Indianapolis-Anderson-Columbus
35	Iowa
36	Jacksonville
37	Kansas City-Overland Park-Kansas City, KS part
38	Kansas City-Overland Park-Kansas City, MO part
39	Lake Charles-Jennings
40	Laredo
41	Las Vegas-Paradise-Pahrump
42	Los Angeles-Long Beach-Riverside
43	Louisville/Jefferson County-Elizabethtown-Scottsburg
44	Maine
45	Memphis
46	Miami-Fort Lauderdale-Pompano Beach
47	Milwaukee-Racine-Waukesha
48	Minneapolis-St. Paul-St. Cloud
49	Mississippi
50	Mobile-Daphne-Fairhope
51	Montana
52	Nashville-Davidson-Murfreesboro-Columbia
53	Nebraska
54	New Hampshire
55	New Mexico
56	New Orleans-Metairie-Bogalusa
57	New York-Newark-Bridgeport, CT part
58	New York-Newark-Bridgeport, NJ part
59	New York-Newark-Bridgeport, NY part
60	North Dakota
61	Oklahoma City-Shawnee
62	Orlando-Deltona-Daytona Beach
63	Philadelphia-Camden-Vineland, NJ part
64	Philadelphia-Camden-Vineland, PA part
65	Phoenix-Mesa-Scottsdale
66	Pittsburgh-New Castle

Table A.1 – continued from previous page

index	CFS Area
67	Portland-Vancouver-Beaverton
68	Raleigh-Durham-Cary
69	Remainder of Alabama
70	Remainder of Arizona
71	Remainder of California
72	Remainder of Colorado
73	Remainder of Connecticut
74	Remainder of Florida
75	Remainder of Georgia
76	Remainder of Hawaii
77	Remainder of Illinois
78	Remainder of Indiana
79	Remainder of Kansas
80	Remainder of Kentucky
81	Remainder of Louisiana
82	Remainder of Maryland
83	Remainder of Massachusetts
84	Remainder of Michigan
85	Remainder of Minnesota
86	Remainder of Missouri
87	Remainder of Nevada
88	Remainder of New Jersey
89	Remainder of New York
90	Remainder of North Carolina
91	Remainder of Ohio
92	Remainder of Oklahoma
93	Remainder of Oregon
94	Remainder of Pennsylvania
95	Remainder of South Carolina
96	Remainder of Tennessee
97	Remainder of Texas
98	Remainder of Utah
99	Remainder of Virginia
100	Remainder of Washington

Table A.1 – continued from previous page

index	CFS Area
101	Remainder of Wisconsin
102	Richmond
103	Rochester-Batavia-Seneca Falls
104	Sacramento-Arden-Arcade-Yuba City
105	Salt Lake City-Ogden-Clearfield
106	San Antonio
107	San Diego-Carlsbad-San Marcos
108	San Jose-San Francisco-Oakland
109	Savannah-Hinesville-Fort Stewart
110	Seattle-Tacoma-Olympia
111	South Dakota
112	St. Louis-St. Charles-Farmington, IL part
113	St. Louis-St. Charles-Farmington, MO part
114	Tampa-St. Petersburg-Clearwater
115	Tucson
116	Tulsa-Bartlesville
117	Vermont
118	Virginia Beach-Norfolk-Newport News
119	Washington-Arlington-Alexandria, DC part
120	Washington-Arlington-Alexandria, MD part
121	Washington-Baltimore-Northern Virginia
122	West Virginia
123	Wyoming

Table A.2: Power law exponents for USA food flow network. B_u and B_d indicated undirected and directed betweenness centrality, respectively. Note that 'Cereal' is for 'cereal grains', 'OthAg' is for 'other agricultural products', 'Animal' is for 'animal feed and products of animal origin, nec', 'Meat' is for 'meat, fish, seafood, and their preparations', and 'Other' is for 'other prepared foodstuffs and fats and oils'.

	s_{in} vs k_{in}	s_{out} vs k_{out}	B_u vs k	B_d vs k
Aggregate	1.335	1.555	2.299	2.300
Cereal	2.308	1.726	2.014	2.018
OthAg	1.253	1.281	2.471	1.805
Animal	1.512	1.175	2.321	2.162
Meat	1.307	1.506	2.166	2.020
Other	1.250	1.425	2.224	2.115

Table A.3: Pearson correlation coefficient (τ) for degree between pairs of linked nodes by direction. Note that the subscript indicates directional pairs, such that 'ii' indicates in-degree, in-degree connection, 'io' indicates in-degree, out-degree connections, 'oo' indicates out-degree, out-degree connections, and 'oi' indicates out-degree, in-degree connections. All other acronyms follow those in Table A.2.

	knn_{ii}	knn_{io}	knn_{oo}	knn_{oi}	knn^W_{ii}	knn^W_{io}	knn^W_{oo}	knn^W_{oi}
Aggregate	-0.128	0.008	0.014	-0.177	0.235	0.416	0.210	0.125
Cereal	0.457	0.467	0.486	0.460	0.485	0.493	0.484	0.506
OthAg	0.092	0.302	0.357	0.419	0.183	0.310	0.439	0.434
Animal	0.293	0.276	0.397	0.416	0.357	0.390	0.491	0.460
Meat	0.074	0.284	0.327	0.309	0.172	0.387	0.487	0.392
Staff	0.027	0.179	0.276	0.216	0.233	0.321	0.437	0.383

Table A.4: Average clustering coefficient for each commodity network. All other acronyms follow those in Table A.2.

	C_{out}	C_{in}	C_{cyc}	C_{mid}	C^W_{out}	C^W_{in}	C^W_{cyc}	C^W_{mid}
Aggregate	0.721	0.736	0.247	0.280	0.780	0.817	0.288	0.321
Cereal	0.048	0.063	0.002	0.024	0.051	0.064	0.000	0.029
OthAg	0.172	0.347	0.036	0.073	0.192	0.366	0.047	0.081
Animal	0.197	0.403	0.056	0.098	0.226	0.438	0.065	0.117
Meat	0.372	0.608	0.093	0.153	0.422	0.679	0.107	0.182
Other	0.489	0.556	0.146	0.185	0.539	0.627	0.180	0.216

Table A.5: Node degree rankings in 2007. Top ten positions according to node in-degree (k_{in}) and out-degree (k_{out}).

rank	k_{in}		k_{out}	
1	Los Angeles-Long Beach-Riverside	86	Remainder of Wisconsin	94
2	Chicago-Naperville-Michigan City	83	Chicago-Naperville-Michigan City	88
3	Atlanta-Sandy Springs-Gainesville	81	Iowa	88
4	Remainder of Pennsylvania	78	Remainder of California	82
5	Dallas-Fort Worth	75	Los Angeles-Long Beach-Riverside	80
6	New York-Newark-Bridgeport, NJ part	68	Remainder of Pennsylvania	79
7	Remainder of Florida	66	New York-Newark-Bridgeport	77
8	Remainder of Wisconsin	66	Remainder of Illinois	75
9	New York-Newark-Bridgeport, NY part	64	Remainder of Indiana	72
10	Remainder of Illinois	64	Arkansas	70

A.1.3 Supplementary Figures

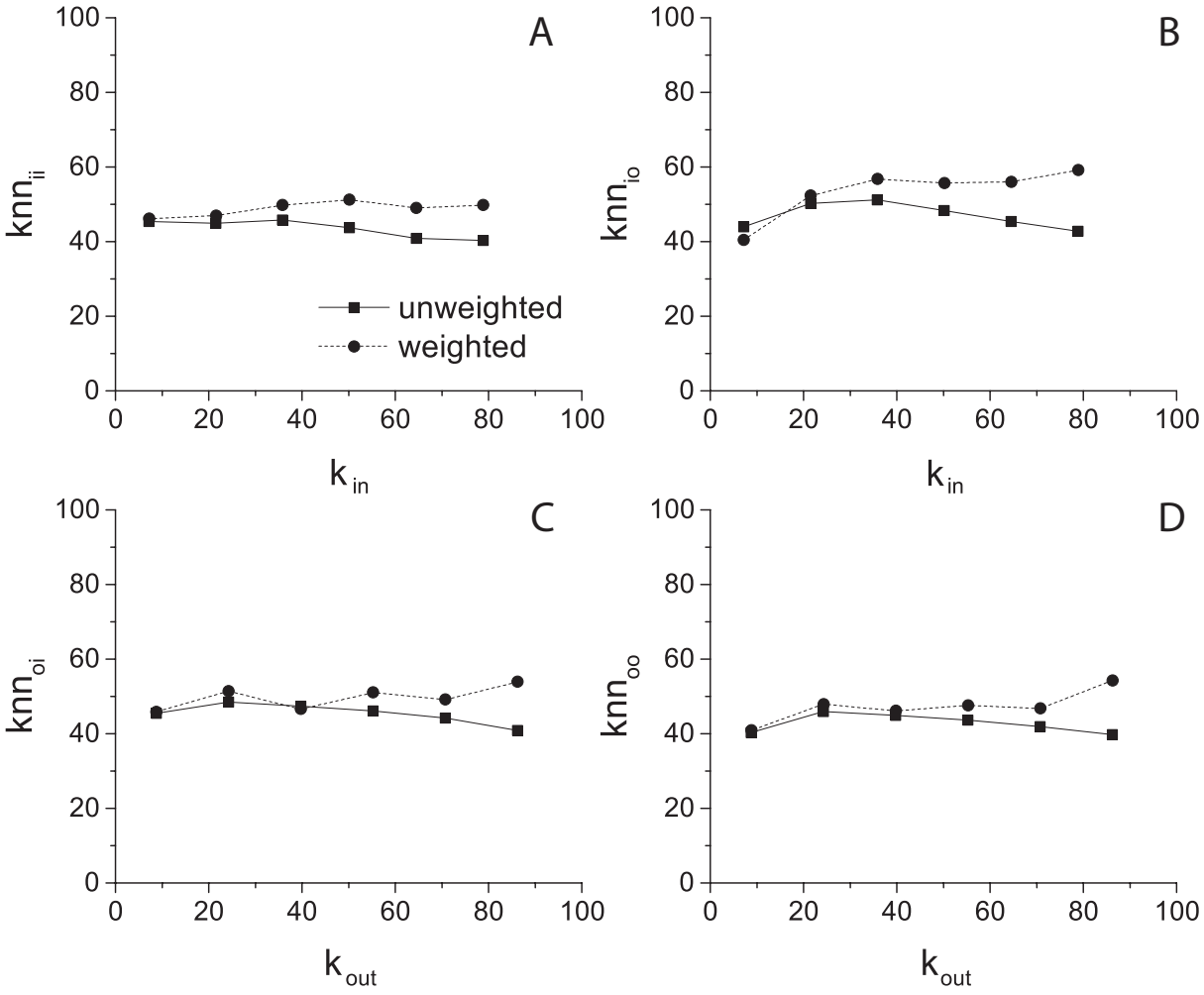

Figure A.1: Nearest neighbor degree (knn) by direction for food flows in the USA. Note that the unweighted measures of knn are unassortative, i.e. uncorrelated with node degree (k), and that the difference between unweighted and weighted knn is small, demonstrating lack of the weighted rich club

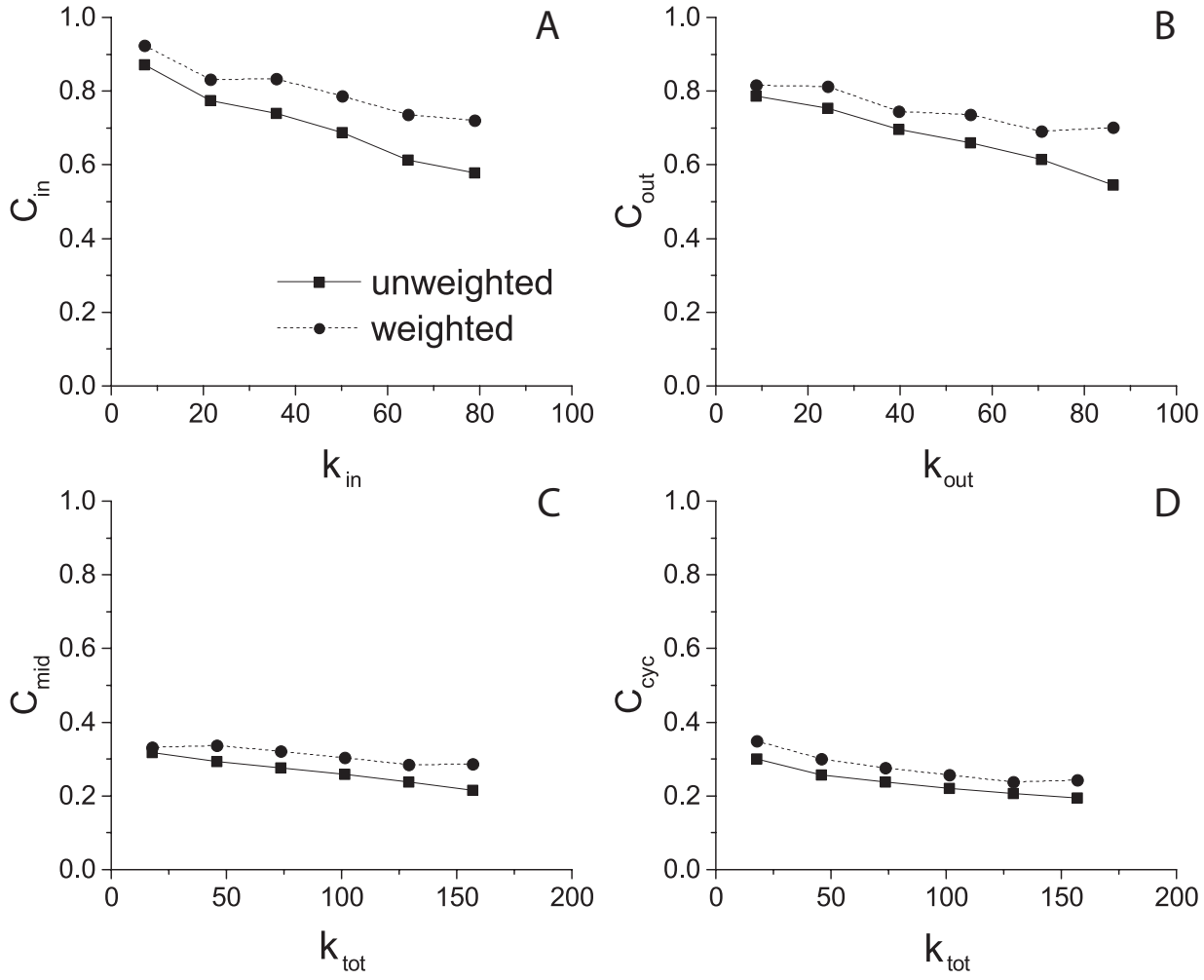

Figure A.2: Clustering by direction for food flows in the USA. Note that the addition of weights to the clustering does not dramatically change the relationship to node degree (k).

A.2 Figure and tables of chapter 2

Table A.6: Summary statistics for USA food flow network. Note that 'Cereal' is for 'cereal grains', 'OthAg' is for 'other agricultural products', 'Animal' is for 'animal feed and products of animal origin, nec', 'Meat' is for 'meat, fish, seafood, and their preparations', and 'Other' is for 'other prepared foodstuffs and fats and oils'. $\overline{k}$ and $\overline{s}$ indicate mean k and mean s, respectively. Values of s are in 10^6 tons.

	Links	Nodes	$\overline{k}$	k_{max}	k_{min}	s_{tot}	$\overline{s}$	s_{max}	s_{min}
Aggregate	8,396	123	68.3	171	4	829.3	6.7	44.8	0.06
Cereal	282	123	2.3	18	0	155.6	1.3	35.5	0
OthAg	1,114	123	9.1	42	0	78.4	0.6	7.7	0
Animal	1,128	123	9.2	48	0	108.3	0.9	15.1	0
Meat	2,630	123	21.4	79	0	74.3	0.6	4.2	0
Other	3,802	123	30.9	115	0	328.0	2.7	15.9	0

Table A.7: Node strength rankings in 2007. Top ten positions according to node in-strength (s_{in}) and out-strength (s_{out}). Note that volume data is provided in 10^6 tons

rank	s_{in}		s_{out}	
1	New Orleans-Metairi-Bogalusa	43.7	Iowa	31.6
2	Remainder of Texas	18.1	Remainder of Illinois	28.3
3	Los Angeles-Long Beach-Riverside	17.1	Remainder of Missouri	20.8
4	Chicago-Naperville-Michigan City	13.4	Nebraska	18.1
5	Remainder of Pennsylvania	12.1	Remainder of California	17.0
6	Remainder of Illinois	12.0	Los Angeles-Long Beach-Riverside	11.9
7	Remainder of California	10.4	Remainder of Pennsylvania	11.4
8	Iowa	10.1	Remainder of Minnesota	11.4
9	Atlanta-Sandy Springs-Gainesville	8.6	Remainder of Wisconsin	10.9
10	Remainder of Louisiana	8.1	Remainder of Indiana	10.4

Table A.8: Node betweenness centrality rankings in 2007. Top ten positions according to node undirected betweenness centrality (B_u) and directed betweenness centrality (B_d). Node betweenness centrality measures the centrality of each node in terms of its location within the global network architecture.

rank	B_u		B_d	
1	Iowa	0.020	Los Angeles-Long Beach-Riverside	0.095
2	Remainder of Illinois	0.013	Chicago-Naperville-Michigan City	0.085
3	Remainder of Missouri	0.010	Remainder of Texas	0.078
4	Nebraska	0.007	Remainder of Pennsylvania	0.074
5	Remainder of California	0.020	New York-Newark-Bridgeport	0.057
6	Los Angeles-Long Beach-Riverside	0.031	Iowa	0.056
7	Remainder of Pennsylvania	0.023	Remainder of California	0.053
8	Remainder of Minnesota	0.007	Remainder of Wisconsin	0.053
9	Remainder of Wisconsin	0.019	Atlanta-Sandy Springs-Gainesville	0.057
10	Remainder of Indiana	0.011	San Jose-San Francisco-Oakland	0.045

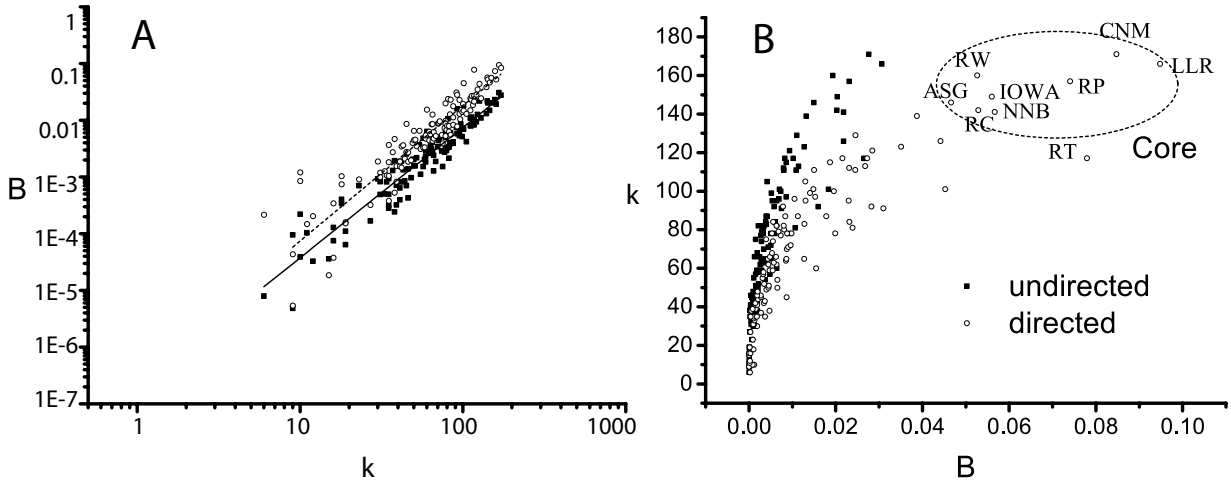

Figure A.3: Relationship between node betweenness centrality (B) and node degree (k) for food flows in the USA. (Panel A) Both undirected and directed B display a power law relationship with K. (Panel B) A core group of nodes is evident for directed B. The core nodes are: 'Los Angeles-Long Beach-Riverside' (LLR), 'Chicago-Naperville-Michigan City' (CNM), 'Remainder of Texas' (RT), 'Remainder of Pennsylvania' (RP), 'New York-Newark-Bridgeport' (NNB), 'Iowa' (IOWA), 'Remainder of California' (RC), 'Remainder of Wisconsin' (RW), and 'Atlanta-Sandy Springs-Gainesville' (ASG).

A.3 Figure of Chapter 3

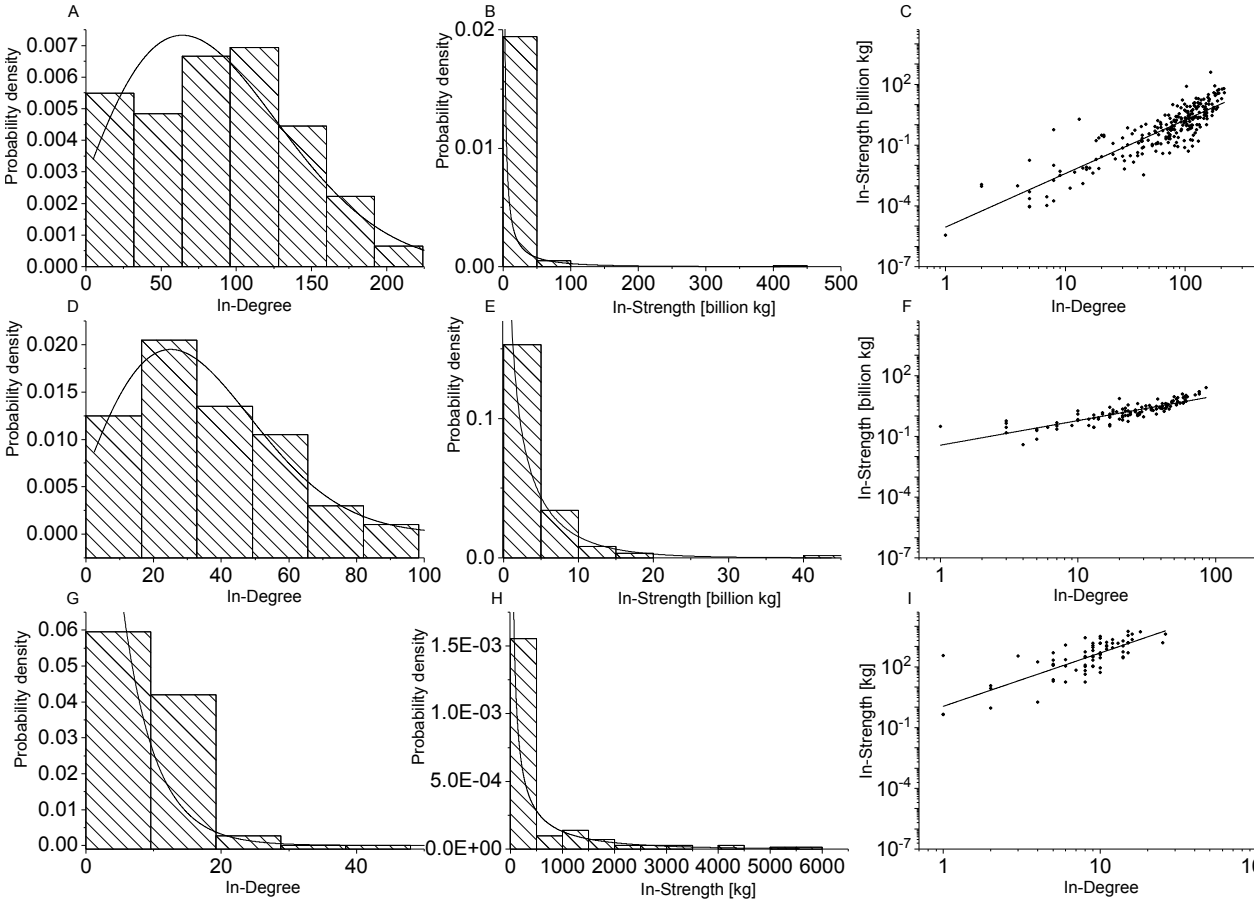

Figure A.4: Network properties for import food flow networks. Global scale is shown in the top row (Panels A, B, C), national scale is shown in the middle row (Panels D, E, F), and village scale is shown in the bottom row (Panels G, H, I). Node in-degree distributions with generalized exponential distributions fit to the data are shown in the first column (Panel A, D, G), node in-strength [kg] distributions with gamma distributions fit to the data are shown in the second column (Panels B, E, H), and power law relationships for node in-strength versus in-degree are shown in the third column (Panels C, F, I).

A.4 Supplementary Materials of Chapter 4

Explanation of degree distribution with three independent random Poisson processes

One potential explanation is that, for any location i, its in-degree is decided by three factors. The first factor is how much food deficit location i has given current connections. When a new inflow link is created, it could be beneficial by reducing the food deficit ('effective') or 'ineffective' as it only passed by. If a link happens to be 'effective', it would reduce food surplus in location i's counterpart equally. By reducing food deficit and food surplus, the motivation for these locations to form a new connection is reduced accordingly. We explain this mechanism with a Poisson process $N_{12}(degree), degree >= 0$. N_{12} causes a potential 'damage' to both location i and its counter part by increasing the hazard rate for failure of both by s_i. There are other factors that could impact location i's new food import link, however, it is impossible to explicitly list them all. These factors might be trading policy, marginal new inflow link benefit, transportation and infrastructure, and etc. For example, in the investigated Alaska villages, households are living far away from the rest of people and transportation is difficult. So it would soon reach a limit that a household can or willing to directly connect. It is more reasonable to exchange as much food (in terms of types and quantity) as possible with a very small amount of households instead of commuting among many households. The marginal benefit of new inflow link soon becomes below zero. These factors are complex and we don't have information. Without prior knowledge, we assume at each degree, both location i and its counterpart have two independent Poisson processes with a fixed failure intensity, λ_i and λ_j. So the hazard rate for location i is $d_i N_i(degree) + s_i N_{12}(degree), d_i = \infty$ (*Ryu*, 1993), suggesting that once location i is not 'willing' to form an inflow link, no new link will be created; as failures from N_{12} cumulate (more food received), it will increase the failure rate (increase the chance of stopping forming new links). Based on this hazard function, we can derive the same probability density function as generalized exponential distribution observed in the empirical degree distribution. Vice versa for food exportation in location i. This stochastic process with three independent Poisson processes can potentially be used to explain the formation mechanism leading to the observed degree distribution. It is interesting to observe that the shared component has greater coefficient than individual in both country and FAF food flow network, while it is the opposite for households in villages. This suggests link formation between countries and FAF areas is more impacted by the trade benefit, while it is more restricted by other factors (e.g., transportation, etc.) among households in Alaska.

REFERENCES

Agarwal, S., and S. Kalla (1996), A generalized gamma distribution and its application in reliabilty, *Communications in Statistics-Theory and Methods*, *25*(1), 201–210.

Aldaya, M. M., A. K. Chapagain, A. Y. Hoekstra, and M. M. Mekonnen (2012), *The water footprint assessment manual: Setting the global standard*, Routledge.

Artin, E. (2015), *The gamma function*, Courier Dover Publications.

Baggio, J. A., S. B. BurnSilver, A. Arenas, J. S. Magdanz, G. P. Kofinas, and M. De Domenico (2016), Multiplex social ecological network analysis reveals how social changes affect community robustness more than resource depletion, *Proceedings of the National Academy of Sciences*, *113*(48), 13,708–13,713.

Balakrishnan, N., A. Basu, and H. Nagaraja (1998), The exponential distribution: Theory, methods and applications, *SIAM Review*, *40*(1), 167–167.

Barabasi, A.-L. (2002), *Linked: The new science of networks*, Perseus Publishing.

Barabási, A.-L., and R. Albert (1997), Emergence of scaling in random networks, *Science*, *286*, 509–512.

Barigozzi, M., G. Fagiolo, and D. Garlaschelli (2010), Multinetwork of international trade: A commodity-specific analysis, *Physical Review E*, *81*, 046,104.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani (2004), The architecture of complex weighted networks, *Proceedings of the National Academy of Sciences*, *101*(11), 3747–3752.

Barthelemy, M. (2004), Betweenness centrality in large complex networks, *The European Physical Journal B*, *38*(2), 163–168.

Berkoff, J. (2003), China: the south–north water transfer projectis it justified?, *Water policy*, *5*(1), 1–28.

Bettencourt, L. M., J. Lobo, D. Helbing, C. Kühnert, and G. B. West (2007), Growth, innovation, scaling, and the pace of life in cities, *Proceedings of the national academy of sciences*, *104*(17), 7301–7306.

Bhattacharya, K., G. Mukherjee, and S. Manna (2007), The international trade network, *Econophysics of markets and business networks*, pp. 139–147.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modelling: a review, *Hydrological processes*, *9*(3-4), 251–290.

Bohannon, R. W. (1995), Standing balance, lower extremity muscle strength, and walking performance of patients referred for physical therapy, *Perceptual and motor skills*, *80*(2), 379–385.

Bouwman, L., K. Klein Goldewijk, K. W. V. D. Hoek, A. H. W. Beusen, D. P. V. Vuuren, J. Willems, M. C. Rufino, and E. Stehfest (2013), Exploring global changes in nitrogen and phosphorus cycles in agriculture induced by livestock production over the 19002050 period, *PNAS*, *110*(52), 20,882–20,887, doi:10.1073/pnas.1012878108.

Burger, M., F. Van Oort, and G.-J. Linders (2009), On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation, *Spatial Economic Analysis*, *4*(2), 167–190.

Burgess, R., and D. Donaldson (2010), Can openness mitigate the effects of weather shocks? Evidence from India's famine era, *American Economic Review: Papers and Proceedings*, pp. 449–453.

Burton, J., B. Eggleston, J. Brenner, A. Truchil, B. A. Zulkiewicz, and M. A. Lewis (2017), Community-based health education programs designed to improve clinical measures are unlikely to reduce short-term costs or utilization without additional features targeting these outcomes, *Population health management*, *20*(2), 93–98.

Census (2017), U.s. census bureau.

CFS (2013), Commodity Flow Survey, $http://www.census.gov/econ/cfs/$.

Chini, C. M., M. Konar, and A. S. Stillwell (2017), Direct and indirect urban water footprints of the united states, *Water Resources Research*, *53*(1), 316–327.

CIA (2017), Central intelligence agency.

COMTRADE (2016), U.n. comtrade database.

Costa, L., F. Rodrigues, G. Travieso, and P. V. Boas (2007), Characterization of complex networks: a survey of measurements, *Advances in Physics*, *56*(1), 167–242.

Cuéllar, A. D., and M. E. Webber (2010), Wasted Food, Wasted Energy: The Embedded Energy in Food Waste in the United States, *Environmental Science & Technology*, *44*(16), 64646469, doi:10.1021/es100310d.

Cukier, R., C. Fortuin, K. E. Shuler, A. Petschek, and J. Schaibly (1973), Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory, *The Journal of chemical physics*, *59*(8), 3873–3878.

Dalin, C., M. Konar, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe (2012), Evolution of the global virtual water trade network, *Proc. Nat. Acad. Sci.*, *109*(16), 5989–5994, doi:10.1073/pnas.1203176109.

Damgaard, C., and J. Weiner (2000), Describing inequality in plant size or fecundity, *Ecology*, *81*, 1139–1142.

Dang, Q., X. Lin, and M. Konar (2015), Agricultural virtual water flows within the united states, *Water Resources Research*, *51*(2), 973–986.

D'Odorico, P., J. A. Carr, F. Laio, L. Ridolfi, and S. Vandoni (2014), Feeding humanity through global food trade, *Earth's Future*, *2*(9), 458–469.

Ercsey-Ravasz, M., Z. Toroczkai, Z. Lakner, and J. Baranyi (2012), Complexity of the international agro-food trade network and its impact on food safety, *PLoS ONE*, *7*(5), e37,810, doi:10.1371/journal.pone.0037810.

Erdos, P., and A. Rényi (1960), On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17–60.

Fagiolo, G. (2007), Clustering in complex directed networks, *Phys. Rev. E*, *76*, 026,107.

Fagiolo, G., J. Reyes, and S. Schiavo (2008), On the topological properties of the world trade web: A weighted network analysis, *Physica A*, *387*, 3868–3873.

FAO (2013), Staple foods: What do people eat?, `http://www.fao.org/docrep/u8480e/U8480E07.htm#Staple\%20foods\%20What\%20do\%20people\%20eat`.

Faraway, J. J. (2016), *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Chapman and Hall/CRC.

Feed America (2018a), Understanding hunger and food insecurity.

Feed America (2018b), Hunger a harsh reality for 14 million children nationwide.

Foley, J. A., R. DeFries, G. P. Asner, C. Barfor, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder (2005), Global Consequences of Land Use, *Science*, *309*.

Foundation, W. A. (2018), Hunger statistics.

Fricke, D., K. Finger, and T. Lux (2013), On assortative and disassortative mixing scale-free networks: The case of interbank credit network, *Kiel Working Paper No. 1830*.

Garlaschelli, D., and M. I. Loffredo (2005), Structure and evolution of the world trade network, *Physica A*, *355*, 491–499.

Gini, C. (1909), Concentration and dependency ratios (in italian), english translation in, *Rivista di Politica Economica*, *87*, 769–789.

Godfray, H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin (2010), Food security: The challenge of feeding 9 billion people, *Science*, *327*(5967), 812–818, doi:10.1126/science.1185383.

Guimera, R., A. Diaz-Guilera, F. Giralt, and A. Arenas (2006), The real communication network behind the formal chart: Community structure in organizations, *Journal of Economic Behavior and Organization*, *61*, 653–667.

Gupta, R. D., and D. Kundu (1999), Theory & methods: Generalized exponential distributions, *Australian & New Zealand Journal of Statistics*, *41*(2), 173–188.

Han, S., and N. Soroka (2014), Us trade overview, 2013, *Change*, *6*, 7.

Hanley, J. A., and B. J. McNeil (1982), The meaning and use of the area under a receiver operating characteristic (roc) curve., *Radiology*, *143*(1), 29–36.

Hansen, M. C., S. V. Stehman, and P. V. Potapov (2010), Quantification of global gross forest cover loss, *PNAS*, *107*(19), 8650–8655, doi:10.1073/pnas.0912668107.

Hertel, T., M. Burke, and D. Lobell (2010), The poverty implications of climate-induced crop yield changes by 2030, *Global Environmental Change*, *20*(4), 577–585.

Hoekstra, A. Y., and M. M. Mekonnen (2012), The water footprint of humanity, *Proceedings of the national academy of sciences*, *109*(9), 3232–3237.

Holland, P. W., and R. E. Welsch (1977), Robust regression using iteratively reweighted least-squares, *Communications in Statistics-theory and Methods*, *6*(9), 813–827.

Holt-Giménez, E., A. Shattuck, M. Altieri, H. Herren, and S. Gliessman (2012), We already grow enough food for 10 billion people and still can't end hunger.

Hoover, E. (1941), Interstate redistribution of population, 1850-1941, *J. Econ. Hist.*, *1*, 199–205.

Isserman, A. M., and J. Westervelt (2006), 1.5 million missing numbers: Overcoming employment suppression in county business patterns data, *International Regional Science Review*, doi:10.1177/0160017606290359.

Jackson, M. (2008), *Social and economic networks*, Princeton University Press.

Jackson, M. O. (2010), *Social and economic networks*, Princeton university press.

Jackson, M. O., and A. Watts (2001), The existence of pairwise stable networks.

Jackson, M. O., T. Rodriguez-Barraquer, and X. Tan (2012), Social capital and social quilts: Network patterns of favor exchange, *American Economic Review*, *102*(5), 1857–97.

Jeong, H., Z. Néda, and A.-L. Barabási (2003), Measuring preferential attachment in evolving networks, *EPL (Europhysics Letters)*, *61*(4), 567.

Kalapala, V., V. Sanwalani, A. Clauset, and C. Moore (2006), Scale invariance in road networks, *Phys. Rev. E*, *73*(2).

Kaluza, P., A. Kolzsch, M. T. Gastner, and B. Blasius (2010), The complex network of global cargo ship movements, *Journal of the Royal Society Interface*, *7*, 1093–1103, doi: 10.1098/rsif.2009.0495.

Kastner, T., M. Rivas, W. Koch, and S. Nonhebel (2012), Global changes in diets and the consequences for land requirements for food, *Proc. Nat. Acad. Sci.*, *109*(18), 6868–6872, doi:10.1073/pnas.1117054109.

Kohavi, R., et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada.

Konar, M., and K. Caylor (2013), Virtual water trade and development in Africa, *Hydrol. Earth Syst. Sci.*, *17*, 3969–3982, doi:10.5194/hess-17-3969-2013.

Konar, M., C. Dalin, S. Suweis, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe (2011), Water for food: The global virtual water trade network, *Water Resources Research*, *47*(W05520), doi:10.1029/2010WR010307.

Konar, M., C. Dalin, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe (2012), Temporal dynamics of blue and green virtual water trade networks, *Water Resources Research*, *48*(W07509), doi:10.1029/2012WR011959.

Konar, M., T. P. Evans, M. Levy, C. A. Scott, T. J. Troy, C. J. Vörösmarty, and M. Sivapalan (2016a), Water resources sustainability in a globalizing world: who uses the water?, *Hydrological Processes*, *30*(18), 3330–3336.

Konar, M., J. J. Reimer, Z. Hussein, and N. Hanasaki (2016b), The water footprint of staple crop trade under climate and policy scenarios, *Environmental Research Letters*, *11*(3), 035,006.

Konar, M., X. Lin, B. Ruddell, and M. Sivapalan (2018), Scaling properties of food flow networks, *PLoS ONE*, *13*(7), e0199,498, doi:10.1371/journal.pone.0199498.

Kyriakopoulos, F., S. Thurner, C. Puhr, and S. W. Schmitz (2009), Network and eigenvalue analysis of financial transaction networks, *The European Physical Journal B*, *71*, 523–531.

Leroy, V., B. B. Cambazoglu, and F. Bonchi (2010), Cold start link prediction, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 393–402, ACM.

Levin, S. A. (1992), The problem of pattern and scale in ecology: the robert h. macarthur award lecture, *Ecology*, *73*(6), 1943–1967.

Liang, S., H. Wang, S. Qu, T. Feng, D. Guan, H. Fang, and M. Xu (2016), Socioeconomic Drivers of Greenhouse Gas Emissions in the United States, *Environmental Science & Technology*, *50*, 7535–7545, doi:10.1021/acs.est.6b00872.

Liang, X.-Z., Y. Wu, R. G. Chambers, D. L. Schmoldt, W. Gao, C. Liu, Y.-A. Liu, C. Sun, and J. A. Kennedy (2017), Determining climate effects on US total agricultural productivity, *Proc. Nat. Acad. Sci.*, *114*(12), E2285–E229, doi:10.1073/pnas.1615922114.

Lin, X., Q. Dang, and M. Konar (In Review), A network analysis of food flows within the USA, *Environ. Sci. and Tech.*

Liu, J., V. Hull, M. Batistella, R. DeFries, T. Dietz, F. Fu, T. W. Hertel, R. C. Izaurralde, E. F. Lambin, S. Li, et al. (2013), Framing sustainability in a telecoupled world, *Ecology and Society*, *18*(2).

Lobell, D. B., W. Schlenker, and J. Costa-Roberts (2011), Climate trends and global crop production since 1980, *Science*, *333*(6042), 616–620, doi:10.1126/science.1204531.

Long, S. P., A. Marshall-Colon, and X.-G. Zhu (2015), Meeting the global food demand of the future by engineering crop photosynthesis and yield potential, *Cell*, *161*(1), 56–66, doi:10.1016/j.cell.2015.03.019.

Lü, L., and T. Zhou (2011), Link prediction in complex networks: A survey, *Physica A: statistical mechanics and its applications*, *390*(6), 1150–1170.

MacDonald, G. K., E. M. Bennett, and S. R. Carpenter (2012), Embodied phosphorus and the global connections of united states agriculture, *Environmental Research Letters*, *7*(4), 044,024.

Marston, L., and M. Konar (2017), Drought impacts to water footprints and virtual water transfers of the c entral v alley of c alifornia, *Water Resources Research*, *53*(7), 5756–5773.

Marston, L., M. Konar, X. Cai, and T. J. Troy (2015), Virtual groundwater transfers from overexploited aquifers in the united states, *Proceedings of the National Academy of Sciences*, *112*(28), 8561–8566.

Marston, L., Y. Ao, M. Konar, M. Mekonnen, and A. Y. Hoekstra (2018), High-Resolution Water Footprints of Production of the United States, *Water Resources Research*, *54*(3), doi:10.1002/2017WR021923.

Mas, D., A. Vignes, and G. Weisbuch (2007), Networks and syndication strategies: Does a venture capitalist need to be in the center?, *ERMES-CNRS Universite Pantheon Assas Paris II, Working Paper No. 07-14*.

Masucci, A., D. Smith, A. Crooks, and M. Batty (2009), Random planar graphs and the london street network, *UCL Working Paper Series*, *146*.

May, R., S. Levin, and G. Sugihara (2008), Ecology for bankers, *Nature*, *451*(21), 893–895.

Melissa, M., et al. (2017), Foodâ energyâ water nexus: Quantifying embodied energy and ghg emissions from irrigation through virtual water transfers in food trade, *ACS sustainable chemistry*.

Micklin, P. P. (1988), Desiccation of the aral sea: a water management disaster in the soviet union, *Science*, *241*(4870), 1170–1176.

Miguens, J., and J. Mendes (2008), Weighted and directed network on traveling patterns, *Biowire*, *5151*, 145–154.

Milo, R., S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon (2004), Superfamilies of evolved and designed networks, *Science*, *303*(5663), 1538–1542, doi:10.1126/science.1089167.

Morris, M. D. (1991), Factorial sampling plans for preliminary computational experiments, *Technometrics*, *33*(2), 161–174.

Motter, A. E., and Y.-C. Lai (2002), Cascade-based attacks on complex networks, *Physical Review E*, *66*, 065,102, doi:10.1103/PhysRevE.66.065102.

Nesheim, M., M. Oria, and P. Yih (Eds.) (2015), *A Framework for Assessing Effects of the Food System*, National Academies Press, doi:10.17226/18846.

Newman, M. (2001a), Scientific collaboration networks: 1. Network construction and fundamental results, *Physical Review E*, *64*, 016,131.

Newman, M. (2002), Assortative mixing in networks, *Phys. Rev. Lett.*, *89*, 208,701.

Newman, M. (2004), Coauthorship networks and patterns of scientific collaborations, *Proc. Nat. Acad. Sci.*, *101*, 5200–5205.

Newman, M., A.-L. Barabasi, and D. J. Watts (2006), *The structure and dynamics of networks*, 1st ed., Princeton University Press.

Newman, M. E. (2001b), Clustering and preferential attachment in growing networks, *Physical review E*, *64*(2), 025,102.

Nolin, D. A. (2010), Food-sharing networks in lamalera, indonesia, *Human Nature*, *21*(3), 243–268.

Oak Ridge National Laboratory (2011), County-to-county distance matrix, `http://cta.ornl.gov/transnet/SkimTree.htm`.

Oak Ridge National Laboratory (2015), Freight Analysis Framework version 4, `http://faf.ornl.gov/fafweb/`.

O'Bannon, C., J. Carr, D. Seekell, and P. D'Odorico (2013), Globalization of agricultural pollution due to international trade, *Hydrol. Earth Syst. Sci. Discuss.*, *10*, 11,221–11,239, doi:10.5194/hessd-10-11221-2013.

Pennock, D. M., G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles (2002), Winners don't take all:Characterizing the competition for links on the web, *Proc. Nat. Acad. Sci.*, *99*(8), 5207–5211.

Peters, G., J. Minx, C. Weber, and O. Edenhofer (2011), Growth in emission transfers via international trade from 1990 to 2008, *Proc. Nat. Acad. Sci.*, *108*(21), 8903–8908, doi:10.1073/pnas.1006388108.

Porkka, M., M. Kummu, S. Siebert, and O. Varisl (2013), From food insufficiency towards trade dependency: A historical analysis of global food availability, *PLoS ONE*, *8*(12), e82,714, doi:10.1371/journal.pone.0082714.

Puma, M. J., S. Bose, S. Y. Chon, and B. I. Cook (2015), Assessing the evolving fragility of the global food system, *Environmental Research Letters*, *10*(2), 024,007.

python (2018), python, `http://geocoder.readthedocs.io`.

Ramsey, F., and D. Schafer (2012), *The statistical sleuth: a course in methods of data analysis*, Cengage Learning.

Reimer, J. J., and M. Li (2010), Trade costs and the gains from trade in crop agriculture, *American Journal of Agricultural Economics*, *92*(4), 1024–1039.

Rinaldi, S. M., J. P. Peerenboom, and T. K. Kelly (2001), Identifying, understanding, and analyzing critical infrastructure interdependencies, *IEEE Control Systems*, *21*(6), 11–25.

Rushforth, R. R., and B. L. Ruddell (2016), The vulnerability and resilience of a city's water footprint: The case of flagstaff, arizona, usa, *Water Resources Research*, *52*(4), 2698–2714.

Rushforth, R. R., and B. L. Ruddell (2018), A Spatially Detailed and Economically Complete Blue Water Footprint of the United States, *Hydrology and Earth System Sciences*, doi: 10.5194/hess-2017-650.

Ruxton, G. D. (2006), The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test, *Behavioral Ecology*, *17*(4), 688–690.

Ryu, K. (1993), An extension of marshall and olkin's bivariate exponential distribution, *Journal of the American Statistical Association*, *88*(424), 1458–1465.

Salako, F., and G. Tian (2003), Soil water depletion under various leguminous cover crops in the derived savanna of west africa, *Agriculture, Ecosystems & Environment*, *100*(2-3), 173–180.

Saltelli, A., S. Tarantola, and K.-S. Chan (1999), A quantitative model-independent method for global sensitivity analysis of model output, *Technometrics*, *41*(1), 39–56.

Sayles, J. S., and J. A. Baggio (2017), Social–ecological network analysis of scale mismatches in estuary watershed restoration, *Proceedings of the National Academy of Sciences*, *114*(10), E1776–E1785.

Schipanski, M., and E. Bennett (2012), The influence of agricultural trade and livestock production on the global phosphorus cycle, *Ecosystems*, *15*(2), 256–268, doi: 10.1007/s10021-011-9507-x.

SCTG (2017), 2012 commodity flow survey standard classification of transported goods (sctg).

Seekell, D., P. D'Odorico, and M. Pace (2011), Virtual water transfers unlikely to redress inequality in global water use, *Environ. Res. Lett.*, *6*(024017).

Serrano, M., and M. Boguna (2003), Topology of the world trade web, *Phys. Rev. E*, *68*(1), 015,101.

Seto, K. C., A. Reenberg, C. G. Boone, M. Fragkias, D. Haase, T. Langanke, P. Marcotullio, D. K. Munroe, B. Olah, and D. Simon (2012), Urban land teleconnections and sustainability, *Proceedings of the National Academy of Sciences*, *109*(20), 7687–7692.

Shen, Q., S. Liang, M. Konar, Z. Zhu, A. Chiu, X. Jia, and M. Xu (2018), Virtual water scarcity risk to the global trade system, *Environmental Science & Technology*, *52*(2), 673–683, doi:10.1021/acs.est.7b04309.

Shutters, S. T., and R. Muneepeerakul (2012), Agricultural trade networks and patterns of economic development, *PLoS ONE*, *7*(7), e39,756, doi:10.1371/journal.pone.0039756.

Smith, T. M., A. L. Goodkind, T. Kim, R. E. Pelton, K. Suh, and J. Schmitt (2017), Subnational mobility and consumption-based environmental accounting of us corn in animal protein and ethanol supply chains, *Proceedings of the National Academy of Sciences*, *114*(38), E7891–E7899.

Sobol, I. M. (2001), Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, *Mathematics and computers in simulation*, *55*(1-3), 271–280.

Thai, M. T., and P. Pardalos (Eds.) (2012), *Handbook of optimization in complex networks*, 544 pp., Springer.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tilman, D. (1999), Global environmental impacts of agricultural expansion: The need for sustainable and efficient practices, *PNAS*, *96*(11), 5995–6000, doi:10.1073/pnas.96.11.5995.

Tinbergen, J. (1962), *An Analysis of World Trade Flows, in: Shaping the world economy*, Twentieth Century Fund.

US Bureau of Economic Analysis (2014), Input-Output Accounts Data, `https://www.bea.gov/industry/io_annual.htm`.

US Bureau of Economic Analysis (2017), Local area personal income and employment, `https://www.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=7#reqid=70&step=1&isuri=1`.

US Census Bureau (2015a), 2012 Commodity Flow Survey Public Use Microdata, `https://www.census.gov/econ/cfs/pums.html`.

US Census Bureau (2015b), 2012 Economic Census, `http://www.census.gov/data.html`.

US Census Bureau (2018), Us census bureau usa trade database, `https://usatrade.census.gov/index.php`.

US Department of Agriculture (2014), National agricultural statistics service quick stats, `http://quickstats.nass.usda.gov`.

USDA (2013), United states department of agriculture economic research service, `http://www.ers.usda.gov/data-products.aspx`.

Venkatramanan, S., S. Wu, and etc. (2017), Towards robust models of food flows and their role in invasive species spread, in *Big Data (Big Data), 2017 IEEE International Conference on*, pp. 435–444, IEEE.

Vitousek, P., H. Mooney, J. Lubchenko, and J. Melillo (1997), Human domination of Earth's ecosystems, *Science*, *277*(5325), 494–499.

Wang, R., J. B. Zimmerman, C. Wang, D. F. Vivanco, and E. G. Hertwich (2017), Freshwater Vulnerability beyond Local Water Stress: Heterogeneous Effects of Water-Electricity Nexus Across the Continental United States, *Environmental Science & Technology*, *51*(17), 98999910, doi:10.1021/acs.est.7b01942.

Wasserman, S., and K. Faust (1994), *Social network analysis: Methods and applications*, 1st ed., Cambridge University Press.

Watts, D. (1999), *Small worlds: the dynamics of networks between order and randomness*, 1st ed., Princeton University Press.

Watts, D. J., and S. H. Strogatz (1998), Collective dynamics of small-world networks, *Nature*, *393*(6684), 440–442.

Weber, C. L., and H. S. Matthews (2008), Food-Miles and the Relative Climate Impacts of Food Choices in the United States, *Environmental Science & Technology*, *42*(10), 35083513, doi:10.1021/es702969f.

West, G. B., J. H. Brown, and B. J. Enquist (1997), A general model for the origin of allometric scaling laws in biology, *Science*, *276*(5309), 122–126.

Wu, F., and H. Guclu (2013), Global maize trade and food security: Implications from a social network model, *Risk Analysis*, *33*(12), doi:10.1111/risa.12064.

Xu, M., B. R. Allenby, and J. C. Crittenden (2011), Interconnectedness and resilience of the U.S. economy, *Advances in Complex Systems*, *14*(5), 649–672, doi:10.1142/S0219525911003335.

Xu, M., M. Weissburg, J. P. Newell, and J. C. Crittenden (2012), Developing a science of infrastructure ecology for sustainable urban systems.